

Organization **TC1600**

Bidg./Room **REMSEN**

U. S. DEPARTMENT OF COMMERCE

COMMISSIONER FOR PATENTS

P.O. BOX 1450

ALEXANDRIA, VA 22313-1450

IF UNDELIVERABLE RETURN IN TEN DAYS

OFFICIAL BUSINESS

AN EQUAL OPPORTUNITY EMPLOYER

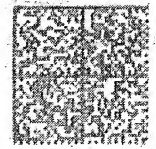
7590  
ANNE MARIE K  
KNOBE MARTE  
620 NEWPORT CE  
SIXTEENTH FLOC  
NEWPORT BEACH

**KNOB**  
KNOB620 926605006 1B03 21 11/20/04  
FORWARD TIME EXP ETON TO SEND  
KNOBE MARTE STE 1400  
2040 MAIN ST STE 1400  
IRVINE CA 92614-8214  
RETURN TO SENDER

RECEIVED

NOV 24 2004

TECH CENTER 1600/2900



UNITED STATES POSTAGE  
\$02.21  
NOV 16 2004  
MAILED FROM ZIP CODE 22202



# UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE  
United States Patent and Trademark Office  
Address: COMMISSIONER FOR PATENTS  
P.O. Box 1450  
Alexandria, Virginia 22313-1450  
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/063,617	05/03/2002	Dan L. Eaton	P3230R1C001-168	4531

7590 11/16/2004

ANNE MARIE KAISER  
KNOBBE MARTENS OLSON & BEAR  
620 NEWPORT CENTER DRIVE  
SIXTEENTH FLOOR  
NEWPORT BEACH, CA 92660

EXAMINER

ROMEO, DAVID S

ART UNIT	PAPER NUMBER
----------	--------------

1647

DATE MAILED: 11/16/2004

Please find below and/or attached an Office communication concerning this application or proceeding.

# Office Action Summary

Application No.

10/063,617

Applicant(s)

EATON ET AL.

Examiner

David S Romeo

Art Unit

1647

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

## Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If the period for reply specified above is less than thirty (30) days, a reply within the statutory minimum of thirty (30) days will be considered timely.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

## Status

- 1) ☒ Responsive to communication(s) filed on 03 May 2002.
- 2a) ☐ This action is FINAL. 2b) ☒ This action is non-final.
- 3) ☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

## Disposition of Claims

- 4) ☒ Claim(s) 1-13 is/are pending in the application.
- 4a) Of the above claim(s) \_\_\_\_\_ is/are withdrawn from consideration.
- 5) ☐ Claim(s) \_\_\_\_\_ is/are allowed.
- 6) ☒ Claim(s) 1-13 is/are rejected.
- 7) ☐ Claim(s) \_\_\_\_\_ is/are objected to.
- 8) ☐ Claim(s) \_\_\_\_\_ are subject to restriction and/or election requirement.

RECEIVED

NOV 24 2004

TECH CENTER 1600/2900

## Application Papers

- 9) ☒ The specification is objected to by the Examiner.
- 10) ☐ The drawing(s) filed on \_\_\_\_\_ is/are: a) ☐ accepted or b) ☐ objected to by the Examiner.  
Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).  
Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
- 11) ☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

## Priority under 35 U.S.C. § 119

- 12) ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
- a) ☐ All b) ☐ Some \* c) ☐ None of:
- ☐ Certified copies of the priority documents have been received.
  - ☐ Certified copies of the priority documents have been received in Application No. \_\_\_\_\_.
  - ☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

\* See the attached detailed Office action for a list of the certified copies not received.

## Attachment(s)

- ☒ Notice of References Cited (PTO-892)
- ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
- ☒ Information Disclosure Statement(s) (PTO-1449 or PTO/SB/08)  
Paper No(s)/Mail Date 0902.
- ☐ Interview Summary (PTO-413)  
Paper No(s)/Mail Date. \_\_\_\_\_.
- ☐ Notice of Informal Patent Application (PTO-152)
- ☐ Other: \_\_\_\_\_.

### DETAILED ACTION

The preliminary amendment filed 09/10/2002 has been entered. Claims 1-13 are pending and being examined.

5

#### *Specification*

The disclosure is objected to because it contains an embedded hyperlink and/or other form of browser-executable code. Applicant is required to delete the embedded hyperlink and/or other form of browser-executable code. See MPEP § 608.01.

10

#### *Claim Rejections - 35 USC §§ 101, 112*

35 U.S.C. 101 reads as follows:

Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title.

15

The following is a quotation of the first paragraph of 35 U.S.C. 112:

The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same and shall set forth the best mode contemplated by the inventor of carrying out his invention.

20

The following is a quotation of the second paragraph of 35 U.S.C. 112:

The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

25

Claims 1-13 are rejected under 35 U.S.C. 101 because the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility.



Art Unit: 1647

The present claims are drawn to or encompass an isolated polypeptide comprising the amino acid sequence of SEQ ID NO: 110 (PRO1753) or comprising an amino acid sequence having a recited % identity thereto.

5 The present specification discloses a nucleotide sequence (SEQ ID NO: 109) of a native sequence PRO1753 cDNA, wherein SEQ ID NO: 109 is a clone designated as "DNA68883-1691" (paragraph 0135). FIG. 110 shows the amino acid sequence (SEQ ID NO: 110) derived from the coding sequence of SEQ ID NO: 109 shown in FIG. 109 (paragraph 0136). The specification discloses uses for PRO polynucleotides and polypeptides in general (paragraphs 0316-0360; pages 86-100). Example 18 (Tumor  
10 Versus Normal Differential Tissue Expression Distribution) discloses that DNA68883-1691 is more highly expressed in esophageal tumor as compared to normal esophagus (page 143).

The present specification discloses that secreted proteins and membrane-bound proteins and receptors have widely varying activities (paragraphs 0002-0004). This  
15 finding establishes that secreted proteins and membrane-bound proteins and receptors have very diverse functions and makes it clear that classification of a protein as a secreted protein or a membrane-bound protein or receptor does not identify it as having a specific function. The specification provides no basis for concluding which, if any, of the varied activities of secreted proteins and membrane-bound proteins and receptors is possessed  
20 by the PRO1753 polypeptide. There is no evidence that a skilled artisan would have appreciated the identification of the PRO1753 polypeptide, without more, would have suggested any specific patentable utility.

Art Unit: 1647

The disclosed uses for PRO polynucleotides and polypeptides in general (paragraphs 0316-0360) are not specific to the PRO1753 polypeptide.

Although the specification discloses that DNA68883-1691 is more highly expressed in esophageal tumor as compared to normal esophagus (page 143), the specification provides no information regarding the absolute values of the differences in transcript levels and provides no information regarding level of expression, activity, or role of the PRO1753 polypeptide in cancer. The art demonstrates that increased transcript levels do not necessarily correlate with increased polypeptide levels. See Haynes (U), who studied more than 80 proteins relatively homogeneous in half-life and expression level, and found no strong correlation between protein and transcript level. For some genes, equivalent mRNA levels translated into protein abundances which varied more than 50-fold. Haynes concluded that the protein levels cannot be accurately predicted from the level of the corresponding mRNA transcript (page 1863, second paragraph, and Figure 1).

Hancock (V) states that "the markers that are generated by proteomics are not always consistent with the markers that are generated from expression profiling" (full paragraph 2).

Therefore, the art indicates that transcript levels are not always correlated with protein levels.

Furthermore, the literature cautions researchers from drawing conclusions based on small changes in transcript expression levels between normal and cancerous tissue. For example, Hu (W) analyzed 2286 genes that showed a greater than 1-fold difference in mean expression level between breast cancer samples and normal samples in a

Art Unit: 1647

microarray (p. 408, middle of right column). Hu discovered that, for genes displaying a 5-fold change or less in tumors compared to normal, there was no evidence of a correlation between altered gene expression and a known role in the disease. However, among genes with a 10-fold or more change in expression level, there was a strong and significant correlation between expression level and a published role in the disease (see discussion section).

In addition, Wang (X) indicates that differential display is the first of many steps required in the discovery of a novel pharmacological target, especially given that the function of the factor is most likely unknown. Therefore, further action should be taken to characterize the functions of a particular gene of interest, including ... validation for the importance of the gene in disease processes. See page 279, column 2, full paragraph 1.

Finally, one skilled in the art recognizes that although structural similarity can serve to classify a protein as related to other known proteins this classification is insufficient to establish a function or biological significance for the protein because ancient duplications and rearrangements of protein-coding segments have resulted in complex gene family relationships. Duplications can be tandem or dispersed and can involve entire coding regions or modules that correspond to folded protein domains. As a result, gene products may acquire new specificities, altered recognition properties, or modified functions. Extreme proliferation of some families within an organism, perhaps at the expense of other families, may correspond to functional innovations during evolution. See Henikoff (Y), page 609, Abstract. Accordingly, one skilled in the art would not accept mere homology as establishing a function of protein because gene products may acquire new specificities, altered recognition properties, or modified

Art Unit: 1647

functions. Rather, homology complements experimental data accumulated for the homologous protein in understanding the homologous protein's biological role.

Although, the presence of a protein module in a protein of interest adds potential insight into its function and guides experiments, insight into the biological function of a protein

5 cannot be automated. However, homology can be used to guide further research. See Henikoff (Y), paragraph bridging pages 613-614, through page 614, paragraph bridging columns 1-2.

Haynes, Hancock, Hu, Wang, and Henikoff are evidence that the specification fails to disclose enough information about the invention to make its usefulness

10 immediately apparent to those familiar with the technological field of the invention. This countervailing evidence shows that the skilled artisan would have a legitimate basis to doubt the utility of the PRO1753 polypeptide. The skilled artisan would not know if PRO1753 polypeptide expression could, should, or would be upregulated, down-regulated, or unchanged in cancer. Therefore, the disclosure that DNA68883-1691 is

15 more highly expressed in esophageal tumor as compared to normal esophagus does not impute a specific, substantial, and credible utility to the PRO1753 polypeptide. Based on the present disclosure, one skilled in the art would be required to carry out further research to identify or reasonably confirm a "real world" context of use. Utilities that require or constitute carrying out further research to identify or reasonably confirm a

20 "real world" context of use are not substantial utilities. Therefore, the increased transcript levels of DNA68883-1691 in esophageal tumor as compared to normal esophagus does not establish a substantial or real-world use for the claimed polypeptide. Thus, the present disclosure is simply a starting point for further research and investigation into

Art Unit: 1647

potential practical uses of the claimed polypeptides. See *Brenner v. Manson*, 148

U.S.P.Q. 689 (Sus. Ct, 1966), wherein the court held that:

5 "The basic quid pro quo contemplated by the Constitution and the  
Congress for granting a patent monopoly is the benefit derived by  
the public from an invention with substantial utility", "[u]nless and  
until a process is refined and developed to this point-where specific  
benefit exists in currently available form-there is insufficient  
10 justification for permitting an applicant to engross what may prove  
to be a broad field", and "a patent is not a hunting license", "[i]t is  
not a reward for the search, but compensation for its successful  
conclusion."

Claims 1-13 are also rejected under 35 U.S.C. 112, first paragraph. Specifically,  
since the claimed invention is not supported by either a specific and substantial asserted  
15 utility or a well established utility for the reasons set forth above, one skilled in the art  
clearly would not know how to use the claimed invention.

Claims 1-5, 12-13 are rejected under 35 U.S.C. 112, first paragraph, as failing to  
comply with the enablement requirement. The claim(s) contains subject matter which  
20 was not described in the specification in such a way as to enable one skilled in the art to  
which it pertains, or with which it is most nearly connected, to make and/or use the  
invention.

The claims are directed to or encompass a polypeptide having at least 80% amino  
acid sequence identity to the polypeptide of SEQ ID NO: 110, to said polypeptide lacking  
25 its associated signal peptide, or to the extracellular domain thereof. The claims are broad  
because they do not require the claimed polypeptide to be identical to the disclosed  
PRO1753 polypeptide and because the claims have no functional limitation.

The first paragraph of 35 U.S.C. 112; that paragraph requires that scope of claims must bear a reasonable correlation to scope of enablement provided by specification to persons of ordinary skill in the art; in cases involving predictable factors, such as mechanical or electrical elements, a single embodiment provides broad enablement in the sense that, once imagined, other embodiments can be made without difficulty and their performance characteristics predicted by resort to known scientific laws; in cases involving unpredictable factors, such as most chemical reactions and physiological activity, scope of enablement varies inversely with degree of unpredictability of factors involved.

10       The PRO1753 polypeptide appears to be a secreted polypeptide. However, the present specification discloses that secreted proteins and membrane-bound proteins and receptors have widely varying activities (paragraphs 0002-0004). This finding establishes that secreted proteins and membrane-bound proteins and receptors have very diverse functions and makes it clear that classification of a protein as a secreted protein or  
15 a membrane-bound protein or receptor does not identify it as having a specific function. The specification provides no basis for concluding which, if any, of the varied activities of secreted proteins and membrane-bound proteins and receptors is possessed by the PRO1753 polypeptide. There is no evidence that a skilled artisan would have appreciated the identification of the PRO1753 polypeptide, without more, would have suggested any  
20 specific use. Therefore, the knowledge that a protein is a secreted polypeptide does not provide predictability about its function.

There are no working examples of polypeptides with an amino acid sequence less than 100% identical to the amino acid sequence of SEQ ID NO: 110. The examiner is

Art Unit: 1647

aware that working examples are not required. Lack of a working example, however, is a factor to be considered, especially in cases involving an unpredictable and undeveloped art.

The specification does not provide guidance for using polypeptides related to (i.e., 80%-99% identity) but not identical to SEQ ID NO: 110. Specifically, the instant specification does not identify those amino acid residues in the amino acid sequence of PRO1753 which are essential for its biological activity and structural integrity and those residues which are either expendable or substitutable. In the absence of this information, the skilled artisan is left to an unduly extensive amount of random, trial and error experimentation wherein a polypeptide comprising the amino acid sequence of SEQ ID NO: 110 is randomly mutated and randomly assayed for a useful activity. Further, there does not appear to be a functionally and structurally analogous protein which has been identified in the prior art for which this information is known and could be extrapolated to the PRO1753 polypeptide by analogy. In any case, while a specification need not disclose what is well known in the art, that rule does not excuse an applicant from providing a complete disclosure. It is the specification, not the knowledge of one skilled in the art, that must supply the novel aspects of an invention in order to constitute adequate enablement. Based on the teachings of the present specification, the skilled artisan would not know how to use such non-identical polypeptides absent undue experimentation. To practice the instant invention in a manner consistent with the breadth of the claims would not require just a repetition of work that is described in the instant application but a substantial inventive contribution on the part of a practitioner which would involve the determination of those amino acid residues in the amino acid

Art Unit: 1647

sequence of SEQ ID NO: 110 which are required for the functional and structural integrity of the PRO1753 polypeptide. It is this additional characterization of that single disclosed, naturally occurring protein that constitutes undue experimentation.

For these reasons, which include the complexity and unpredictability of the nature of the invention and art in terms of the diversity of secreted proteins and membrane-bound proteins and receptors and lack of knowledge about function(s) associated with the PRO1753 polypeptide and its variants, the lack of working examples, the lack of direction or guidance for using polypeptides that are not identical to SEQ ID NO: 110, and the breadth of the claims for structure without function, it would require undue experimentation to use the invention commensurate in scope with the claims.

Claims 1-5, 12, 13 are rejected under 35 U.S.C. 112, first paragraph, as failing to comply with the written description requirement. The claim(s) contains subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention.

The claims are drawn to or encompass a polypeptide having at least 80%, 85%, 90%, 95% or 99% sequence identity with a SEQ ID NO: 110, to said SEQ ID NO: lacking its associated signal peptide, or to the extracellular domain of said SEQ ID NO:.

The claims do not require that the polypeptide possess any particular biological activity, nor any particular conserved structure, or other disclosed distinguishing feature. Thus, the claims are drawn to a genus of polypeptides that is defined only by sequence identity.



To provide adequate written description and evidence of possession of a claimed genus, the specification must provide sufficient distinguishing identifying characteristics of the genus. The factors to be considered include disclosure of complete or partial structure, physical and/or chemical properties, functional characteristics,

5 structure/function correlation, methods of making the claimed product, or any combination thereof. In this case, the only factor present in the claim is a partial structure in the form of a recitation of percent identity. There is not even identification of any particular portion of the structure that must be conserved. Accordingly, in the absence of sufficient recitation of distinguishing identifying characteristics, the specification does  
10 not provide adequate written description of the claimed genus.

Vas-Cath Inc. v. Mahurkar, 19USPQ2d 1111, clearly states “applicant must convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession of the invention. The invention is, for purposes of the 'written description' inquiry, whatever is now claimed.” (See page 1117.) The  
15 specification does not “clearly allow persons of ordinary skill in the art to recognize that [he or she] invented what is claimed.” (See Vas-Cath at page 1116). As discussed above, the skilled artisan cannot envision the detailed chemical structure of the encompassed genus of polypeptides, and therefore conception is not achieved until reduction to practice has occurred, regardless of the complexity or simplicity of the method of  
20 isolation. Adequate written description requires more than a mere statement that it is part of the invention and reference to a potential method of isolating it. The compound itself is required. See *Fiers v. Revel*, 25 USPQ2d 1601 at 1606 (CAFC 1993) and *Amgen Inc. v. Chugai Pharmaceutical Co. Ltd.*, 18 USPQ2d 1016.

Art Unit: 1647

One cannot describe what one has not conceived. See *Fiddes v. Baird*, 30 USPQ2d 1481 at 1483. In *Fiddes*, claims directed to mammalian FGF's were found to be unpatentable due to lack of written description for that broad class. The specification provided only the bovine sequence.

5           Therefore, only isolated polypeptides comprising the amino acid sequence set forth in SEQ ID NO: 110, but not the full breadth of the claim meets the written description provision of 35 U.S.C. §112, first paragraph. Applicant is reminded that *Vas-Cath* makes clear that the written description provision of 35 U.S.C. §112 is severable from its enablement provision (see page 1115).

10

Claims 1-6, 9, 10, 12, 13 are rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

15           The PRO1753 polypeptide is disclosed as a soluble or secreted protein, and is not disclosed as being expressed on a cell surface. Accordingly, the limitation "extracellular domain" is indefinite, as the art does not recognize soluble or secreted proteins as having such domains. Further, if the protein had an extracellular domain, the recitation of "the extracellular domain ... lacking its associated signal sequence" is indefinite as a signal sequence is not generally considered to be part of an extracellular domain, as signal  
20           sequences are cleaved from said domains in the process of secretion from the cell. The metes and bounds are not clearly set forth.

Art Unit: 1647

**Conclusion**

No claims are allowable.

ANY INQUIRY CONCERNING THIS COMMUNICATION OR EARLIER COMMUNICATIONS FROM THE EXAMINER SHOULD BE DIRECTED TO DAVID S. ROMEO WHOSE TELEPHONE NUMBER IS (571) 272-0890. THE EXAMINER CAN NORMALLY BE REACHED ON MONDAY THROUGH FRIDAY FROM 7:30 A.M. TO 4:00 P.M. IF ATTEMPTS TO REACH THE EXAMINER BY TELEPHONE ARE UNSUCCESSFUL, THE EXAMINER'S SUPERVISOR, BRENDA BRUMBACK, CAN BE REACHED ON (571)272-0961.

IF SUBMITTING OFFICIAL CORRESPONDENCE BY FAX, APPLICANTS ARE ENCOURAGED TO SUBMIT OFFICIAL CORRESPONDENCE TO THE FOLLOWING TC 1600 BEFORE AND AFTER FINAL RIGHTFAX NUMBERS:

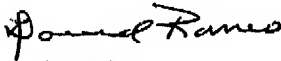
BEFORE FINAL (703) 872-9306

AFTER FINAL (703) 872-9307

CUSTOMERS ARE ALSO ADVISED TO USE CERTIFICATE OF FACSIMILE PROCEDURES WHEN SUBMITTING A REPLY TO A NON-FINAL OR FINAL OFFICE ACTION BY FACSIMILE (SEE 37 CFR 1.6 AND 1.8).

FAXED DRAFT OR INFORMAL COMMUNICATIONS SHOULD BE DIRECTED TO THE EXAMINER AT (571) 273-0890.

ANY INQUIRY OF A GENERAL NATURE OR RELATING TO THE STATUS OF THIS APPLICATION OR PROCEEDING SHOULD BE DIRECTED TO THE GROUP RECEPTIONIST WHOSE TELEPHONE NUMBER IS (703) 308-0196.



DAVID ROMEO  
PRIMARY EXAMINER  
ART UNIT 1647

DSR  
NOVEMBER 9, 2004

FORM PTO-1449

U.S. DEPARTMENT OF COMMERCE  
PATENT AND TRADEMARK OFFICEATTY. DOCKET NO.  
GNE.3230R1C155APPLICATION NO.  
US 10/063,726INFORMATION DISCLOSURE STATEMENT  
BY APPLICANT

(SEE SEVERAL SHEETS IF NECESSARY)

APPLICANT  
Eaton et al.

RECEIVED

FILING DATE  
May 8, 2002GROUP  
1645

SEP 19 2002

## U.S. PATENT DOCUMENTS

TECH CENTER 1600/2900

EXAMINER INITIAL		DOCUMENT NUMBER	DATE	NAME	CLASS	SUBCLASS	FILING DATE (IF APPROPRIATE)
PR	1.	5,536,637	07/16/96	Jacobs			

## FOREIGN PATENT DOCUMENTS

EXAMINER INITIAL		DOCUMENT NUMBER	DATE	COUNTRY	CLASS	SUBCLASS	TRANSLATION	
							YES	NO

## OTHER DOCUMENTS (INCLUDING AUTHOR, TITLE, DATE, PERTINENT PAGES, ETC.)

EXAMINER INITIAL		
PR	2.	Klein et al. Selection for Genes Encoding Secreted Proteins and Receptors. <i>Proc. Natl. Acad. Sci.</i> , 93:7108-7113 (1996)
PR	3.	Database Search, DNA Sequence Alignments [BLASTN 2.2.1 [July-12-2001], NCBI]
PR	4.	Database Search, Protein Sequence Alignments [BLASTN 2.2.1 [July-12-2001], NCBI]

S:\DOCS\AOK\AOK-9407.DOC  
091202

EXAMINER

Donald Rones

DATE CONSIDERED

11/5/4

\*EXAMINER: INITIAL IF CITATION CONSIDERED, WHETHER OR NOT CITATION IS IN CONFORMANCE WITH MPEP 609; DRAW LINE THROUGH CITATION IF NOT IN CONFORMANCE AND NOT CONSIDERED, INCLUDE COPY OF THIS FORM WITH NEXT COMMUNICATION TO APPLICANT.

**Notice of References Cited**

Application/Control No.

10/063,617

Applicant(s)/Patent Under  
Reexamination  
EATON ET AL.

Examiner

David S Romeo

Art Unit

1647

Page 1 of 2

**U.S. PATENT DOCUMENTS**

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
	A	US-			
	B	US-			
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

**FOREIGN PATENT DOCUMENTS**

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

**NON-PATENT DOCUMENTS**

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Haynes et al. Proteome analysis: biological assay or data archive? Electrophoresis. 1998 Aug;19(11):1862-71.
	V	Hancock WS. Do we have enough biomarkers? J Proteome Res. 2004 Jul-Aug;3(4):685.
	W	Hu et al. Analysis of genomic and proteomic data using advanced literature mining. J Proteome Res. 2003 Jul-Aug;2(4):405-12.
	X	Wang et al. mRNA differential display: application in the discovery of novel pharmacological targets. Trends Pharmacol Sci. 1996 Aug;17(8):276-9.

\*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)  
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

**Notice of References Cited**

Application/Control No.

10/063,617

Applicant(s)/Patent Under  
Reexamination  
EATON ET AL.

Examiner

David S Romeo

Art Unit

1647

Page 2 of 2

**U.S. PATENT DOCUMENTS**

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
	A	US-			
	B	US-			
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

**FOREIGN PATENT DOCUMENTS**

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

**NON-PATENT DOCUMENTS**

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	Henikoff et al. Gene families: the taxonomy of protein paralogs and chimeras. Science. 1997 Oct 24;278(5338):609-14.
	V	
	W	
	X	

\*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)  
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

## Review

Paul A. Haynes  
Steven P. Gygi  
Daniel Figgeys  
Ruedi Aebersold

Department of Molecular  
Biotechnology, University of  
Washington, Seattle, WA, USA

## Proteome analysis: Biological assay or data archive?

In this review we examine the current state of proteome analysis. There are three main issues discussed: why it is necessary to study proteomes; how proteomes can be analyzed with current technology; and how proteome analysis can be used to enhance biological research. We conclude that proteome analysis is an essential tool in the understanding of regulated biological systems. Current technology, while still mostly limited to the more abundant proteins, enables the use of proteome analysis both to establish databases of proteins present, and to perform biological assays involving measurement of multiple variables. We believe that the utility of proteome analysis in future biological research will continue to be enhanced by further improvements in analytical technology.

### Contents

1	Introduction .....	1862
2	Rationale for proteome analysis .....	1862
2.1	Correlation between mRNA and protein expression levels .....	1863
2.2	Proteins are dynamically modified and processed .....	1863
2.3	Proteomes are dynamic and reflect the state of a biological system .....	1863
3	Description and assessment of current proteome analysis technology .....	1863
3.1	Technical requirements of proteome technology .....	1863
3.2	2D electrophoresis — mass spectrometry: a common implementation of proteome analysis .....	1864
3.3	Protein identification by LC-MS/MS, capillary LC-MS/MS and CE-MS/MS .....	1865
3.3.1	LC-MS/MS .....	1865
3.3.2	Capillary LC-MS .....	1865
3.3.3	CE-MS/MS .....	1865
3.4	Assessment of 2-DE-MS proteome technology .....	1866
4	Utility of proteome analysis for biological research .....	1868
4.1	The proteome as a database .....	1868
4.2	The proteome as a biological assay .....	1868
5	Concluding remarks .....	1870
6	References .....	1870

### 1 Introduction

A proteome has been defined as the protein complement expressed by the genome of an organism, or, in multicellular organisms, as the protein complement expressed by a tissue or differentiated cell [1]. In the most common implementation of proteome analysis the proteins extracted from the cell or tissue analyzed are separated by high

resolution two-dimensional gel electrophoresis (2-DE), detected in the gel and identified by their amino acid sequence. The ease, sensitivity and speed with which gel-separated proteins can be identified by the use of recently developed mass spectrometric techniques have dramatically increased the interest in proteome technology. One of the most attractive features of such analyses is that complex biological systems can potentially be studied in their entirety, rather than as a multitude of individual components. This makes it far easier to uncover the many complex, and often obscure, relationships between mature gene products in cells. Large-scale proteome characterization projects have been undertaken for a number of different organisms and cell types. Microbial proteome projects currently in progress include, for example: *Saccharomyces cerevisiae* [2], *Salmonella enterica* [3], *Spiroplasma melliferum* [4], *Mycobacterium tuberculosis* [5], *Ochrobactrum anthropi* [6], *Haemophilus influenzae* [7], *Synechocystis* spp. [8], *Escherichia coli* [9], *Rhizobium leguminosarum* [10], and *Dictyostelium discoideum* [11]. Proteome projects underway for tissues of more complex organisms include those for: human bladder squamous cell carcinomas [12], human liver [13], human plasma [13], human keratinocytes [12], human fibroblasts [12], mouse kidney [12], and rat serum [14]. In this manuscript we critically assess the concept of proteome analysis and the technical feasibility of establishing complete proteome maps, and discuss ways in which proteome analysis and biological research intersect.

### 2 Rationale for proteome analysis

The dramatic growth in both the number of genome projects and the speed with which genome sequences are being determined has generated huge amounts of sequence information, for some species even complete genomic sequences ([15–17]). The description of the state of a biological system by the quantitative measurement of system components has long been a primary objective in molecular biology. With recent technical advances including the development of differential display-PCR [18], cDNA microarray and DNA chip technology [19, 20] and serial analysis of gene expression (SAGE) [21, 22], it is now feasible to establish global and quantitative mRNA expression maps of cells and tissues, in which the sequence of all the genes is known, at a speed and sensitivity which is not matched by current

Correspondence: Professor Ruedi Aebersold, Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, WA, 98195, USA (Tel: +206-685-4235; Fax: +206-685-6392; E-mail: ruedi@u.washington.edu)

Abbreviations: CID, collision-induced dissociation; MS/MS, tandem mass spectrometry; SAGE, serial analysis of gene expression

Keywords: Proteome / Two-dimensional polyacrylamide gel electrophoresis / Tandem mass spectrometry

protein analysis technology. Given the long-standing paradigm in biology that DNA synthesizes RNA which synthesizes protein, and the ability to rapidly establish global, quantitative mRNA expression maps, the questions which arise are why technically complex proteome projects should be undertaken and what specific types of information could be expected from proteome projects which cannot be obtained from genomic and transcript profiling projects. We see three main reasons for proteome analysis to become an essential component in the comprehensive analysis of biological systems. (i) Protein expression levels are not predictable from the mRNA expression levels, (ii) proteins are dynamically modified and processed in ways which are not necessarily apparent from the gene sequence, and (iii) proteomes are dynamic and reflect the state of a biological system.

## 2.1 Correlation between mRNA and protein expression levels

Interpretations of quantitative mRNA expression profiles frequently implicitly or explicitly assume that for specific genes the transcript levels are indicative of the levels of protein expression. As part of an ongoing study in our laboratory, we have determined the correlation of expression at the mRNA and protein levels for a population of selected genes in the yeast *Saccharomyces cerevisiae* growing at mid-log phase (S. P. Gygi *et al.*, submitted for publication). mRNA expression levels were calculated from published SAGE frequency tables [22]. Protein expression levels were quantified by metabolic radiolabeling of the yeast proteins, liquid scintillation counting of the protein spots separated by high resolution 2-DE and mass spectrometric identification of the protein(s) migrating to each spot. The selected 80 samples constitute a relatively homogeneous group with respect to predicted half-life and expression level of the protein products. Thus far, we have found a general trend but no strong correlation between protein and transcript levels (Fig. 1). For some genes studied equivalent mRNA transcript levels translated into protein abundances which varied by more than 50-fold. Similarly, equivalent steady-state protein expression levels were maintained by transcript levels varying by as much as 40-fold (S. P. Gygi *et al.*, submitted). These results suggests that even for a population of genes predicted to be relatively homogeneous with respect to protein half-life and gene expression, the protein levels cannot be accurately predicted from the level of the corresponding mRNA transcript.

## 2.2 Proteins are dynamically modified and processed

In the mature, biologically active form many proteins are post-translationally modified by glycosylation, phosphorylation, prenylation, acylation, ubiquitination or one or more of many other modifications [23] and many proteins are only functional if specifically associated or complexed with other molecules, including DNA, RNA, proteins and organic and inorganic cofactors. Frequently, modifications are dynamic and reversible and may alter the precise three-dimensional structure and the state of activity of a protein. Collectively, the state of modification of the proteins which constitute a biological system

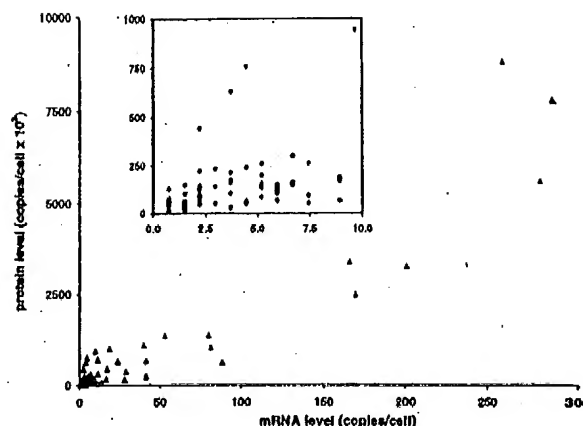


Figure 1. Correlation between mRNA and protein levels in yeast cells. For a selected population of 80 genes, protein levels were measured by  $^{35}\text{S}$ -radiolabeling and mRNA levels were calculated from published SAGE tables. Inset: expanded view of the low abundance region. For more experimental details, also see Figs. 5 and 6, (S. P. Gygi *et al.*, submitted).

are important indicators for the state of the system. The type of protein modification and the sites modified at a specific cellular state can usually not be determined from the gene sequence alone.

## 2.3 Proteomes are dynamic and reflect the state of a biological system

A single genome can give rise to many qualitatively and quantitatively different proteomes. Specific stages of the cell cycle and states of differentiation, responses to growth and nutrient conditions, temperature and stress, and pathological conditions represent cellular states which are characterized by significantly different proteomes. The proteome, in principle, also reflects events that are under translational and post-translational control. It is therefore expected that proteomics will be able to provide the most precise and detailed molecular description of the state of a cell or tissue, provided that the external conditions defining the state are carefully determined. In answer to the question of whether the study of proteomes is necessary for the analysis of biomolecular systems, it is evident that the analysis of mature protein products in cells is essential as there are numerous levels of control of protein synthesis, degradation, processing and modification, which are only apparent by direct protein analysis.

## 3 Description and assessment of current proteome analysis technology

### 3.1 Technical requirements of proteome technology

In biological systems the level of expression as well as the states of modification, processing and macro-molecular association of proteins are controlled and modulated depending on the state of the system. Comprehensive analysis of the identity, quantity and state of modification of proteins therefore requires the detection and



quantitation of the proteins which constitute the system, and analysis of differentially processed forms. There are a number of inherent difficulties in protein analysis which complicate these tasks. First, proteins cannot be amplified. It is possible to produce large amounts of a particular protein by over-expression in specific cell systems. However, since many proteins are dynamically post-translationally modified, they cannot be easily amplified in the form in which they finally function in the biological system. It is frequently difficult to purify from the native source sufficient amounts of a protein for analysis. From a technological point of view this translates into the need for high sensitivity analytical techniques. Second, many proteins are modified and processed post-translationally. Therefore, in addition to the protein identity, the structural basis for differentially modified isoforms also needs to be determined. The distribution of a constant amount of protein over several differentially modified isoforms further reduces the amount of each species available for analysis. The complexity and dynamics of post-translational protein editing thus significantly complicates proteome studies. Third, proteins vary dramatically with respect to their solubility in commonly used solvents. There are few, if any, solvent conditions in which all proteins are soluble and which are also compatible with protein analysis. This makes the development of protein purification methods particularly difficult since both protein purification and solubility have to be achieved under the same conditions. Detergents, in particular sodium dodecyl sulfate (SDS), are frequently added to aqueous solvents to maintain protein solubility. The compatibility with SDS is a big advantage of SDS polyacrylamide gel electrophoresis (SDS-PAGE) over other protein separation techniques. Thus, SDS-PAGE and two-dimensional gel electrophoresis, which also uses SDS and other detergents, are the most general and preferred methods for the purification of small amounts of proteins, provided that activity does not necessarily need to be maintained. Lastly, the number of proteins in a given cell system is typically in the thousands. Any attempt to identify and categorize all of these must use methods which are as rapid as possible to allow completion of the project within a reasonable time frame. Therefore, a successful, general proteomics technology requires high sensitivity, high throughput, the ability to differentiate differentially modified proteins, and the ability to quantitatively display and analyze all the proteins present in a sample.

### 3.2 2-D electrophoresis — mass spectrometry: a common implementation of proteome analysis

The most common currently used implementation of proteome analysis technology is based on the separation of proteins by two-dimensional (IEF/SDS-PAGE) gel electrophoresis and their subsequent identification and analysis by mass spectrometry (MS) or tandem mass spectrometry (MS/MS). In 2-DE, proteins are first separated by isoelectric focusing (IEF) and then by SDS-PAGE, in the second, perpendicular dimension. Separated proteins are visualized at high sensitivity by staining or autoradiography, producing two-dimensional arrays of proteins. 2-DE gels are, at present, the most commonly used means of global display of proteins in complex

samples. The separation of thousands of proteins has been achieved in a single gel [24, 25] and differentially modified proteins are frequently separated. Due to the compatibility of 2-DE with high concentrations of detergents, protein denaturants and other additives promoting protein solubility, the technique is widely used.

The second step of this type of proteome analysis is the identification and analysis of separated proteins. Individual proteins from polyacrylamide gels have traditionally been identified using *N*-terminal sequencing [26, 27], internal peptide sequencing [28, 29], immunoblotting or comigration with known proteins [30]. The recent dramatic growth of large-scale genomic and expressed sequence tag (EST) sequence databases has resulted in a fundamental change in the way proteins are identified by their amino acid sequence. Rather than by the traditional methods described above, protein sequences are now frequently determined by correlating mass spectral or tandem mass spectral data of peptides derived from proteins, with the information contained in sequence databases [31–33].

There are a number of alternative approaches to proteome analysis currently under development. There is considerable interest in developing a proteome analysis strategy which bypasses 2-DE altogether, because it is considered a relatively slow and tedious process, and because of perceived difficulties in extracting proteins from the gel matrix for analysis. However, 2-DE as a starting point for proteome analysis has many advantages compared to other techniques available today. The most significant strengths of the 2-DE-MS approach include the relatively uniform behavior of proteins in gels, the ability to quantify spots and the high resolution and simultaneous display of hundreds to thousands of proteins within a reasonable time frame.

A schematic diagram of a typical procedure of the identification of gel-separated proteins is shown in Fig. 2. Protein spots detected in the gel are enzymatically or chemically fragmented and the peptide fragments are isolated for analysis, as already indicated, most frequently by MS or MS/MS. There are numerous protocols for the generation of peptide fragments from gel-separated proteins. They can be grouped into two categories, digestion in the gel slice [28, 34] or digestion after electrotransfer out of the gel onto a suitable membrane ([29, 35–37] and reviewed in [38]). In most instances either technique is applicable and yields good results. The analysis of MS or MS/MS data is an important step in the whole process because MS instruments can generate an enormous amount of information which cannot easily be managed manually. Recently, a number of groups have developed software systems dedicated to the use of peptide MS and MS/MS spectra for the identification of proteins. Proteins are identified by correlating the information contained in the MS spectra of protein digests or MS/MS spectra of individual peptides with data contained in DNA or protein sequence databases.

The systems we are currently using in our laboratory are based on the separation of the peptides contained in protein digests by narrow bore or capillary liquid chromatog-

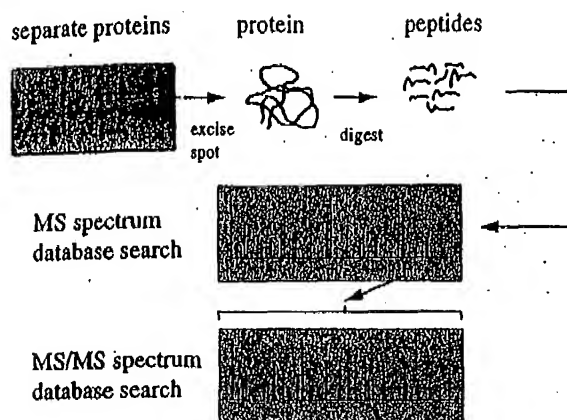


Figure 2. Schematic diagram of a procedure for identification of gel-separated proteins. Peptides can either be separated by a technique such as LC or CE, or infused as a mixture and sorted in the MS. Database searching can either be performed on peptide masses from an MS spectrum, peptide fragment masses from CID spectra of peptides, or a combination of both.

raphy [39, 40] or capillary electrophoresis [41], the analysis of the separated peptides by electrospray ionization (ESI) MS/MS, and the correlation of the generated peptide spectra with sequence databases using the SEQUEST program developed at the University of Washington [32, 33]. The system automatically performs the following operations: a particular peptide ion characterized by its mass-to-charge ratio is selected in the MS out of all the peptide ions present in the system at a particular time; the selected peptide ion is collided in a collision cell with argon (collision-induced dissociation, CID) and the masses of the resulting fragment ions are determined in the second sector of the tandem MS; this experimentally determined CID spectrum is then correlated with the CID spectra predicted from all the peptides in a sequence database which have essentially the same mass as the peptide selected for CID; this correlation matches the isolated peptide with a sequence segment in a database and thus identifies the protein from which the peptide was derived. There are a number of alternative programs which use peptide CID spectra for protein identification, but we use the SEQUEST system because it is currently the most highly automated program and has proven to be successful, versatile and robust.

### 3.3 Protein identification by LC-MS/MS, capillary LC-MS/MS and CE-MS/MS

It has been demonstrated repeatedly that MS has a very high intrinsic sensitivity. For the routine analysis of gel-separated proteins at high sensitivity, the most significant challenge is the handling of small amounts of sample. The crux of the problem is the extraction and transfer of peptide mixtures generated by the digestion of low nanogram amounts of protein, from gels into the MS/MS system without significant loss of sample or introduction of unwanted contaminants. We employ three different systems for introducing gel-purified samples into an MS, depending on the level of sensitivity

required. As an approximate guideline, for samples containing tens of picomoles of peptides, LC-MS/MS is most appropriate; for samples containing low picomole amounts to high femtomole amounts we use capillary LC-MS/MS; and for samples containing femtomoles or less, CE-MS/MS is the method of choice.

#### 3.3.1 LC-MS/MS

The coupling of an MS to an HPLC system using a 0.5 mm diameter or bigger reverse phase (RP) column has been described in detail [42]. This system has several advantages if a large number of samples are to be analyzed and all are available in sufficient quantity. The LC-MS and database searching program can be run in a fully automated mode using an autosampler, thus maximizing sample throughput and minimizing the need for operator interference. The relatively large column is tolerant of high levels of impurities from either gel preparation or sample matrix. Lastly, if configured with a flow-splitter and micro-sprayer [40], analyses can be performed on a small fraction of the sample (less than 5%) while the remainder of the sample is recovered in very pure solvents. This latter feature is particularly useful when an orthogonal technique is also used to analyze peptide fractions, such as scintillation of an introduced radiolabel, and this data can be correlated with peptides identified by CID spectra.

#### 3.3.2 Capillary LC-MS

An increase of sensitivity of approximately tenfold can be achieved by using a capillary LC system with a 100  $\mu$ m ID column rather than a 0.5 mm ID column as referred to above. Since very low flow rates are required for such columns, most reports have used a precolumn flow splitting system for producing solvent gradients. We have recently described the design and construction of a novel gradient mixing system which enables the formation of reproducible gradients at very low flow rates (low nL/min) without the need for flow splitting (A. Ducret *et al.*, submitted for publication). Using this capillary LC-MS/MS system we were able to identify gel-separated proteins if low picomole to high femtomole amounts were loaded onto the gel [40]. This system is as yet not automated and, like all capillary LC systems, is prone to blockage of the columns by microparticulates when analyzing gel-separated proteins.

#### 3.3.3 CE-MS/MS

The highest level of sensitivity for analyzing gel-separated proteins can be achieved by using capillary electrophoresis – mass spectrometry (CE-MS). We have described in the past a solid-phase extraction capillary electrophoresis (SPE-CE) system which was used with triple quadrupole and ion trap ESI-MS/MS systems for the identification of proteins at the low femtomole to sub-femtomole sensitivity level [43, 44]. While this system is highly sensitive, its operation is labor-intensive and its operation has not been automated. In order to devise an analytical system with both the sensitivity of a CE and the level of automation of LC, we have constructed

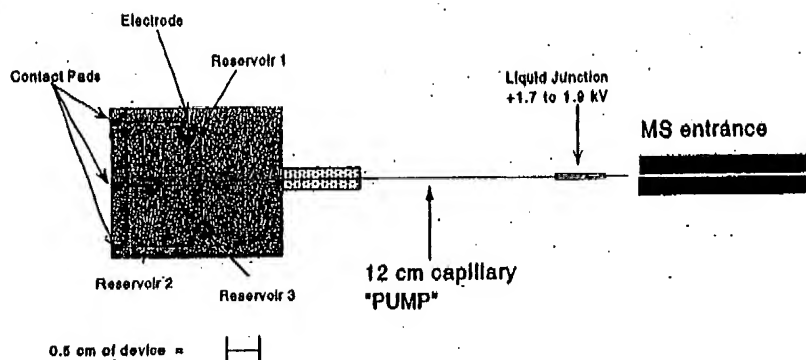


Figure 3. Schematic illustration of a microfabricated analytical system for CE, consisting of a micromachined device, coated capillary electroosmotic pump, and microelectrospray interface. The dimensions of the channels and reservoir are as indicated in the text. The channels on the device were graphically enhanced to make them more visible. Reproduced from [45], with permission.

microfabricated devices for the introduction of samples into ESI-MS for high-sensitivity peptide analysis.

The basic device is a piece of glass into which channels of 10–30  $\mu\text{m}$  in depth and 50–70  $\mu\text{m}$  in diameter are etched by using photolithography/etching techniques similar to the ones used in the semiconductor industry. (A simple device is shown in Fig. 3). The channels are connected to an external high voltage power supply [45]. Samples are manipulated on the device and off the device to the MS by applying different potentials to the reservoirs. This creates a solvent flow by electroosmotic pumping which can be redirected by changing the position of the electrode. Therefore, without the need for valves or gates and without any external pumping, the flow can be redirected by simply switching the position of the electrodes on the device. The direction and rate of the flow can be modulated by the size and the polarity of the electric field applied and also by the charge state of the surface.

The type of data generated by the system is illustrated in Fig. 4, which shows the mass spectrum of a peptide sample representing the tryptic digest of carbonic anhydrase at 290 fmol/ $\mu\text{L}$ . Each numbered peak indicates a peptide successfully identified as being derived from carbonic an-

hydrase. Some of the unassigned signals may be chemical or peptide contaminants. The MS is programmed to automatically select each peak and subject the peptide to CID. The resulting CID spectra are then used to identify the protein by correlation with sequence databases. Therefore, this system allows us to concurrently apply a number of protein digests onto the device, to sequentially mobilize the samples, to automatically generate CID spectra of selected peptide ions and to search sequence databases for protein identification. These steps are performed automatically without the need for user input and proteins can be identified at very low femtomole level sensitivity at a rate of approximately one protein per 15 min.

#### 3.4 Assessment of 2-DE-MS proteome technology

Using a combination of the analytical techniques described above we have identified the 80 protein spots indicated in Fig. 5. The protein pattern was generated by separating a total of 40 microgram of protein contained in a total cell lysate of the yeast strain YPH499 by high resolution 2-DE and silver staining of the separated proteins. To estimate how far this type of proteome analysis can penetrate towards the identification of low abundance proteins, we have calculated the codon bias of the genes encoding the respective proteins. Codon bias is a

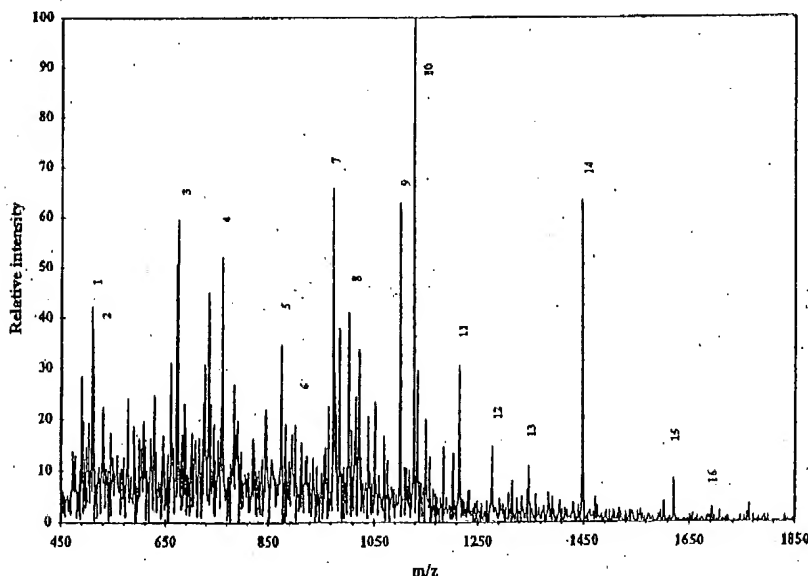


Figure 4. MS spectrum of a tryptic digest of carbonic anhydrase using the microfabricated system shown in Fig. 3. 290 fmol/ $\mu\text{L}$  of carbonic anhydrase tryptic digest was infused into a Finnigan LCQ ion trap MS. Each peak was selected for CID, and those which were identified as containing peptides derived from carbonic anhydrase are numbered. Reproduced from [45], with permission.

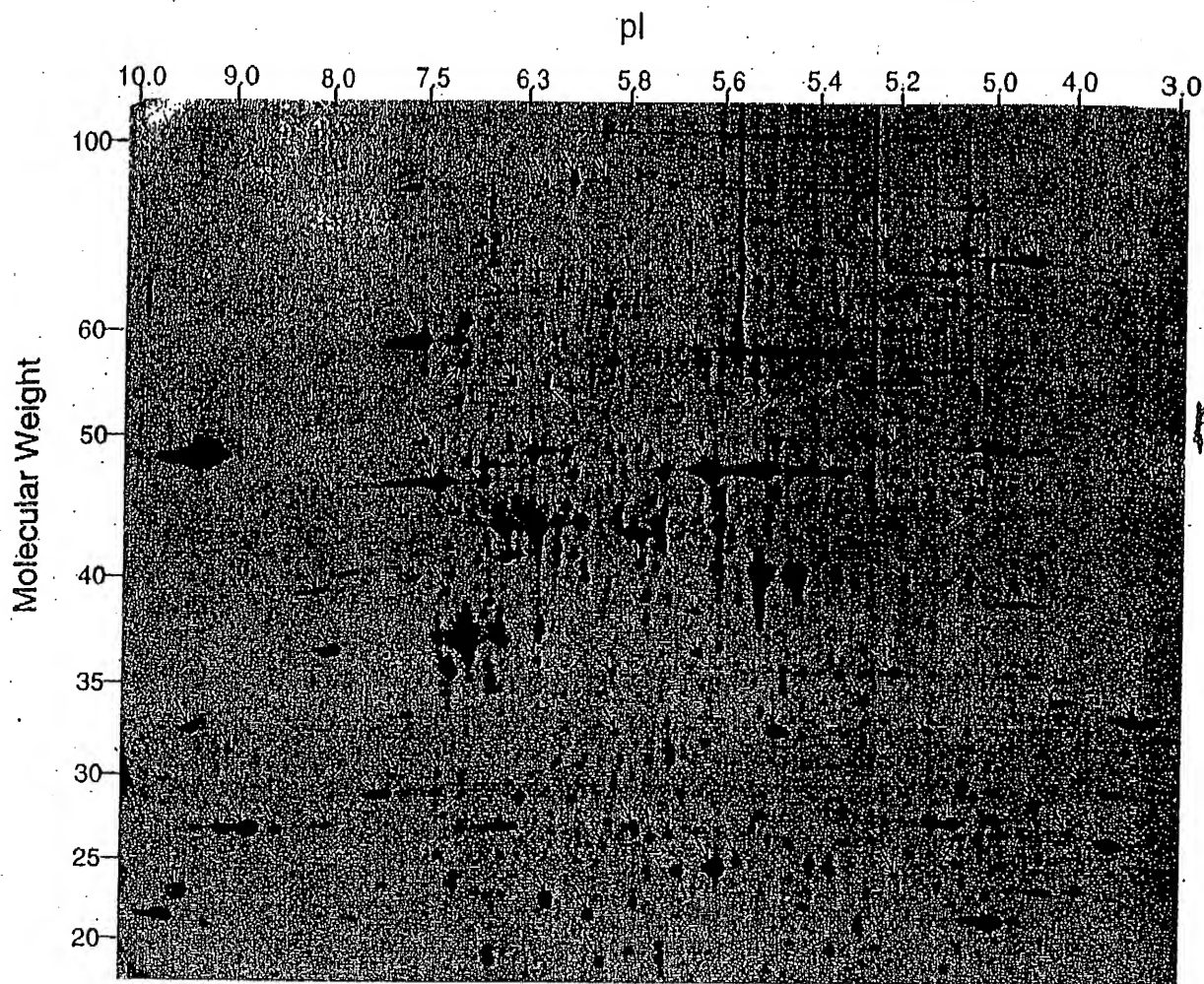


Figure 5. 2-DE separation of a lysate of yeast cells, with identified proteins highlighted. The first dimension of separation was an IPG from pH 3–10, and the second dimension was a 10%T-SDS-PAGE gel. Proteins were visualized by silver staining. Further details of experimental procedures are included in S. P. Gygi *et al.* (submitted).

calculated measure of the degree of redundancy of triplet DNA codons used to produce each amino acid in a particular gene sequence. It has been shown to be a useful indicator of the level of the protein product of a particular gene sequence present in a cell [46]. The general rule which applies is that the higher the value of the codon bias calculated for a gene, the more abundant the protein product of that gene becomes. The calculated codon bias values corresponding to the proteins identified in Fig. 5 are shown in Fig. 6b. Nearly all of the proteins identified (> 95%) have codon bias values of > 0.2, indicating they are highly abundant in cells. In contrast, codon bias values calculated for the entire yeast genome (Fig. 6a) show that the majority of proteins present in the proteome have a codon bias of < 0.2 and are thus of low abundance.

This finding is of considerable importance in our assessment of the current status of proteome analysis technology. It is clear that even using highly sensitive analytical techniques, we are only able to visualize and identify the

more abundant proteins. Since many important regulatory proteins are present only at low abundance, these would not be amenable to analysis using such techniques. This situation would be exacerbated in the analysis of proteomes containing many more proteins than the approximately 6000 gene products present in yeast cells [16]. In the analysis of, for example, the proteome of any human cells, there are potentially 50 000–100 000 gene products [47]. Inherent limitations on the amount of protein that can be loaded on 2-DE, and the number of components that can be resolved, indicate that only the most highly abundant fraction of the many gene products could be successfully analyzed. One approach that has been employed to circumvent these limitations is the use of very narrow range immobilized pH gradient strips for the first-dimension separation of 2-DE [48]. Since only those proteins which focus within the narrow range will enter the second dimension of separation, a much higher sample loading within the desired range is possible. This, in turn, can lead to the visualization and identification of less abundant proteins.



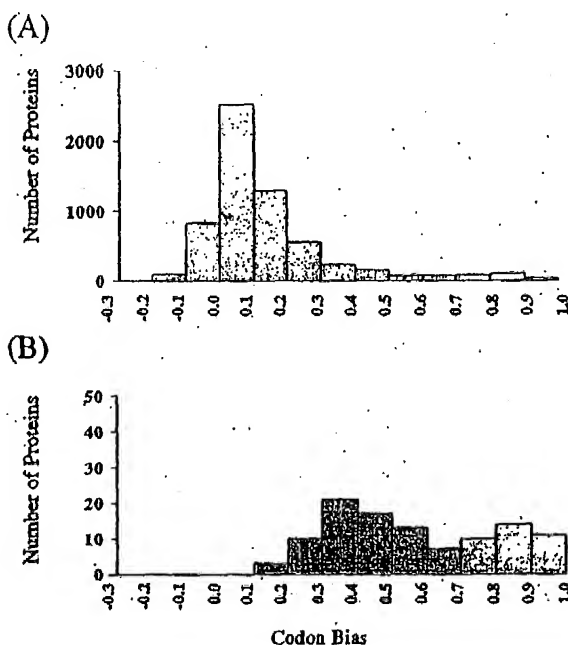


Figure 6. Calculated codon bias values for yeast proteins. (A) Distribution of calculated values for the entire yeast proteome. (B) Distribution of calculated values for the subset of 80 identified proteins also shown in Figs. 1 and 5. Further details of experimental procedures are included in S. P. Gygi *et al.* (submitted).

#### 4 Utility of proteome analysis for biological research

For the success of proteomics as a mainstream approach to the analysis of biological systems it is essential to define how proteome analysis and biological research projects intersect. Without a clear plan for the implementation of proteome-type approaches into biological research projects the full impact of the technology can not be realized. The literature indicates that proteome analysis is used both as a database/data archive, and as a biological assay or biological research tool.

##### 4.1 The proteome as a database

The use of proteomics as a database or data archive essentially entails an attempt to identify all the proteins in a cell or species and to annotate each protein with the known biological information that is relevant for each protein. The level of annotation can, of course, be extensive. The most common implementation of this idea is the separation of proteins by high resolution 2-DE, the identification of each detected protein spot and the annotation of the protein spots in a 2-DE gel database format. This approach is complicated by the fact that it is difficult to precisely define a proteome and to decide which proteome should be represented in the database. In contrast to the genome of a species, which is essentially static, the proteome is highly dynamic. Processes such as differentiation, cell activation and disease can all significantly change the proteome of a species. This is illustrated in Fig. 7. The figure shows two high-resolu-

tion 2-DE maps of proteins isolated from rat serum. Fig. 7A is from the serum of normal rats, while Fig. 7B is from the serum of rats in acute-phase serum after prior treatment with an inflammation-causing agent [49]. It is obvious that the protein patterns are significantly different in several areas, raising the question of exactly which proteome is being described.

Therefore, a comprehensive proteome database of a species or cell type needs to contain all of the parameters which describe the state and the type of the cells from which the proteins were extracted as well as the software tools to search the database with queries which reflect the dynamics of biological systems. A comprehensive proteome database should be capable of quantitatively describing the fate of each protein if specific systems and pathways are activated in the cell. Specifically, the quantity, the degree of modification, the subcellular location and the nature of molecules specifically interacting with a protein as well as the rate of change of these variables should be described. Using these admittedly stringent criteria, there is currently no complete proteome database. A number of such databases are, however, in the process of being constructed. The most advanced among them, in our opinion, are the yeast protein database YPD [50] (accessible at <http://www.ypd.com>) and the human 2D-PAGE databases of the Danish Centre for Human Genome Research [12] (accessible at <http://biobase.dk/cgi-bin/celis>). While neither can be considered complete as not all of the potential gene products are identified, both contain extensive annotation of supplemental information for many of the spots which are positively identified in reference samples.

##### 4.2 The proteome as a biological assay

The use of proteome analysis as a biological assay or research tool represents an alternative approach to integrating biology with proteomics. To investigate the state of a system, samples are subjected to a specific process that allows the quantitative or qualitative measurement of some of the variables which describe the system. In typical biochemical assays one variable (e.g., enzyme activity) of a single component (e.g., a particular enzyme) is measured. Using proteomics as an assay, multiple variables (e.g., expression level, rate of synthesis, phosphorylation state, etc.) are measured concurrently on many (ideally all) of the proteins in a sample. The use of proteomics as an assay is a less far-reaching proposition than the construction of a comprehensive proteome database. It does, however, represent a pragmatic approach which can be adapted to investigate specific systems and pathways, as long as the interpretation of the results takes into account that with current technology not all of the variables which describe the system can be observed (see Section 3.4).

A common implementation of proteome analysis as a biological assay is when a 2-DE protein pattern generated from the analysis of an experimental sample is compared to an array of reference patterns representing different states of the system under investigation. The state of the experimental system at the time the sample was generated is therefore determined by the quantita-

rum.  
7B  
after  
[49].  
ntly  
actly

spe-  
sters  
from  
ware  
flect  
isive  
ively  
tems  
, the  
loca-  
cting  
hese  
edly  
ome  
r, in  
nced  
data-  
and  
entre  
tp://  
con-  
pro-  
ation  
spots  
les.

ty or  
inte-  
state  
cess  
ment  
n. In  
zyme  
r en-  
mul-  
ysis,  
ently  
The  
prop-  
pro-  
natic  
ecific  
on of  
hhol-  
stem

as a  
gener-  
le is  
nting  
The  
mple  
ntita-

tive comparative analysis of hundreds to a few thousand proteins. Comparative analysis of the 2-DE patterns furthermore highlights quantitative and qualitative differences in the protein profiles which correlate with the state of the system. For this type of analysis it is not essential that all the proteins are identified or even visu-

alized, although the results become more informative as more proteins are compared. It is obvious, however, that the possibility to identify any protein deemed characteristic for a particular state dramatically enhances this approach by opening up new avenues for experimentation.

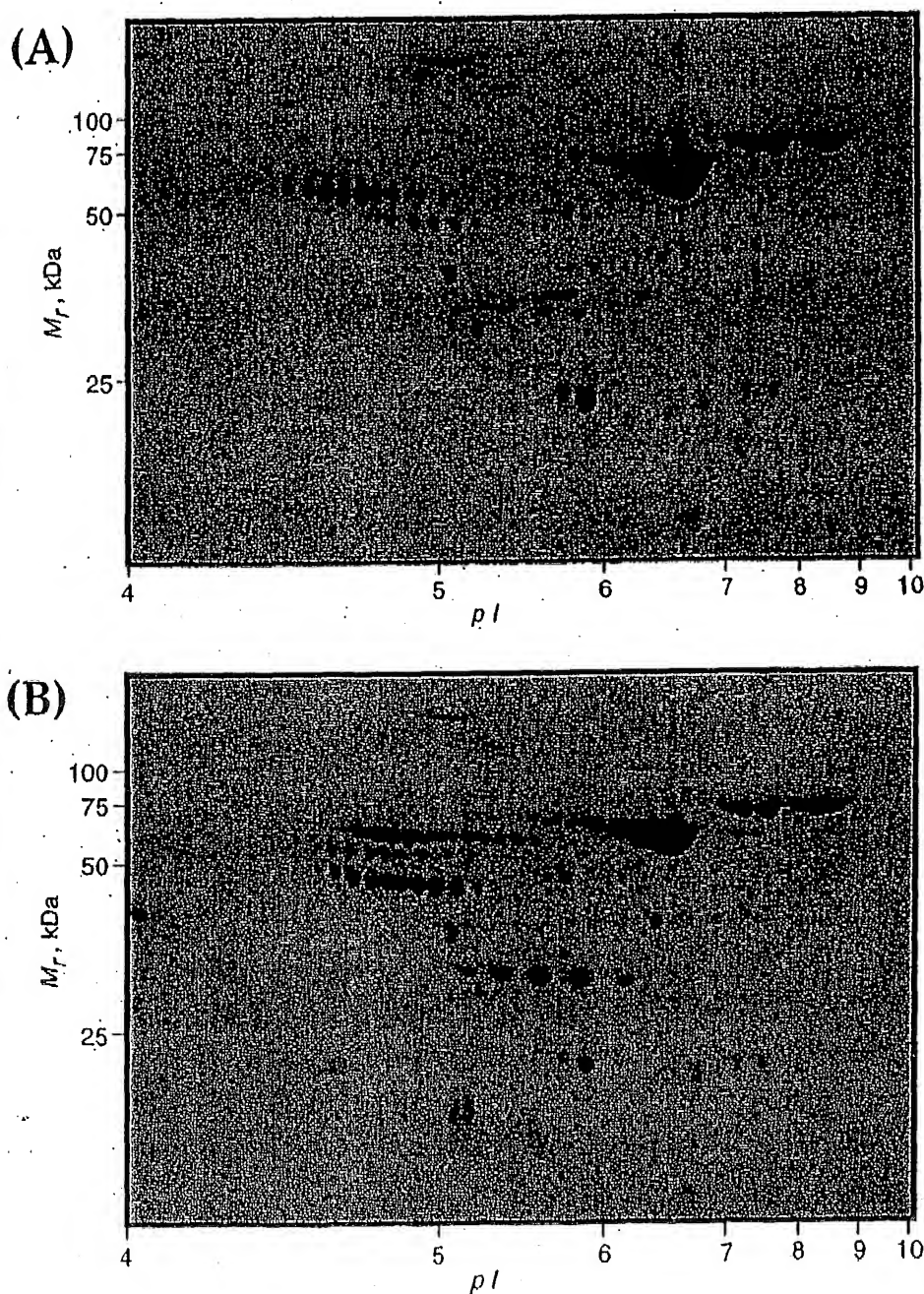


Figure 7. High resolution 2-DE map of proteins isolated from rat serum with or without prior exposure to an inflammation-causing agent. (A) normal rat serum, (B) acute-phase serum from rats which had previously been exposed to an inflammation-causing agent. The first dimension of separation is an IPG from pH 4–10, and the second dimension is a 7.5–17.5%T gradient SDS-PAGE gel. Proteins were visualized by staining with amido black. Further details of experimental procedures are included in [14, 49].

Proteome analysis as a biological assay has been successfully used in the field of toxicology, to characterize disease states or to study differential activation of cells. The approach is limited, of course, by the fact that only the visible protein spots are included in the assay, and it is well known that a substantial but far from complete fraction of cellular proteins are detected if a total cell lysate is separated by 2-DE. Proteins may not be detected in 2-DE gels because they are not abundant enough to be visualized by the detection method used, because they do not migrate within the boundaries (size, pI) resolved by the gel, because they are not soluble under the conditions used, or for other reasons.

A different way to use proteome analysis as a biological assay to define the state of a biological system is to take advantage of the wealth of information contained in 2-DE protein patterns. 2-DE is referred to as two-dimensional because of the electrophoretic mobility and the isoelectric points which define the position of each protein in a 2-DE pattern. In addition to the two dimensions used to generate the protein patterns, a number of additional data dimensions are contained in the protein patterns. Some of these dimensions such as protein expression level, phosphorylation state, subcellular location, association with other proteins, rate of synthesis or degradation indicate the activity state of a protein or a biological system. Comparative analysis of 2-DE protein patterns representing different states is therefore ideally suited for the detection, identification and analysis of suitable markers. Once again it must be emphasized that in this type of experiment only a fraction of the cellular proteins is analyzed. Since many regulatory proteins are of low abundance, this limitation is a concern, particularly in cases in which regulatory pathways are being investigated.

### 5 Concluding remarks

In this report we have addressed three main issues related to proteome analysis. First, we have discussed the rationale for studying proteomes. Second, we have assessed the technical feasibility of analyzing proteomes and described current proteome technology, and third, we have analyzed the utility of proteome analysis for biological research. It is apparent that proteome analysis is an essential tool in the analysis of biological systems. The multi-level control of protein synthesis and degradation in cells means that only the direct analysis of mature protein products can reveal their correct identities, their relevant state of modification and/or association and their amounts. Recently developed methods have enabled the identification of proteins at ever-increasing sensitivity levels and at a high level of automation of the analytical processes. A number of technical challenges, however, remain. While it is currently possible to identify essentially any protein spots that can be visualized by common staining methods, it is apparent that without prior enrichment only a relatively small and highly selected population of long-lived, highly expressed proteins is observed. There are many more proteins in a given cell which are not visualized by such methods. Frequently it is the low abundance proteins that execute key regulatory functions.

We have outlined the two principal ways proteome analysis is currently being used to intersect with biological research projects: the proteome as a database or data archive and proteome analysis as a biological assay. Both approaches have in common that at present they are conceptually and technically limited. Current proteome databases typically are limited to one cell type and one state of a cell and therefore do not account for the dynamics of biological systems. The use of proteome analysis as a biological assay can provide a wealth of information, but it is limited to the proteins detected and is therefore not truly proteome-wide. These limitations in proteomics are to a large extent a reflection of the fact that proteins in their fully processed form cannot easily be amplified and are therefore difficult to isolate in amounts sufficient for analysis or experimentation. The fact that to date no complete proteome has been described further attests to these difficulties. With continued rapid progress in protein analysis technology, however, we anticipate that the goal of complete proteome analysis will eventually become attainable.

*We would like to acknowledge the funding for our work from the National Science Foundation Science and Technology Center for Molecular Biotechnology and from the NIH. We thank Yvan Rochon and Bob Franza for providing the yeast gel shown and Elisabetta Gianazza for providing the rat serum gels shown.*

Received April 21, 1998

### 6 References

- [1] Wilkins, M. R., Pasquall, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J.-C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L., Hochstrasser, D. F., *BioTechnology* 1996, 14, 61-65.
- [2] Hodges, P. E., Payne, W. E., Garrels, J. I., *Nucleic Acids Res.* 1998, 26, 68-72.
- [3] O'Connor, C. D., Farris, M., Fowler, R., Qi, S. Y., *Electrophoresis* 1997, 18, 1483-1490.
- [4] Cordwell, S. J., Basseal, D. J., Humphery-Smith, I., *Electrophoresis* 1997, 18, 1335-1346.
- [5] Urquhart, B. L., Atsalos, T. E., Roach, D., Basseal, D. J., Bjellqvist, B., Britton, W. L., Humphery-Smith, I., *Electrophoresis* 1997, 18, 1384-1392.
- [6] Wasinger, V. C., Bjellqvist, B., Humphery-Smith, I., *Electrophoresis* 1997, 18, 1373-1383.
- [7] Link, A. J., Hays, L. G., Carmack, E. B., Yates III, J. R., *Electrophoresis* 1997, 18, 1314-1334.
- [8] Sazuka, T., Ohara, O., *Electrophoresis* 1997, 18, 1252-1258.
- [9] VanBogelen, R. A., Abshire, K. Z., Moldover, B., Olson, E. R., Neidhardt, F. C., *Electrophoresis* 1997, 18, 1243-1251.
- [10] Guerreiro, N., Redmond, J. W., Rolfe, B. G., Djordjevic, M. A., *Mol. Plant Microbe Interact.* 1997, 10, 506-516.
- [11] Yan, J. X., Tonella, L., Sanchez, J.-C., Wilkins, M. R., Packer, N. H., Gooley, A. A., Hochstrasser, D. F., Williams, K. L., *Electrophoresis* 1997, 18, 491-497.
- [12] Cells, J., Gromov, P., Ostergaard M., Madsen, P., Honoré, B., Dejgaard, K., Olsen, E., Vorum, H., Kristensen, D. B., Gromova, I., Haunso, A., Van Damme, J., Puype, M., Vandekerckhove, J., Rasmussen, H. H., *FEBS Lett.* 1996, 398, 129-134.
- [13] Appel, R. D., Sanchez, J.-C., Bairoch, A., Golaz, O., Miu, M., Vargas, J. R., Hochstrasser, D. F., *Electrophoresis* 1993, 14, 1232-1238.
- [14] Haynes, P., Miller, I., Aebersold, R., Gemeiner, M., Eberlin, I., Lovati, R. M., Manzoni, C., Vignati, M., Gianazza, E., *Electrophoresis* 1998, 19, 1484-1492.

- [15] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-P., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, N. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghegan, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, C. O., Venter, J. C., *Science* 1995, 269, 496-512.
- [16] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S. G., *Science* 1996, 274, 546.
- [17] Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, B. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., Peterson, J., Kerlavage, A. R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M. D., Gocayne, J., Weidman, J., Utterback, T., Wathey, T., McDonald, L., Artlach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M. D., Horst, K., Roberts, K., Hatch, B., Smith, H. O., Venter, J. C., *Nature* 1997, 390, 580-586.
- [18] Lian, P., Pardee, A. B., *Science* 1992, 257, 967-971.
- [19] Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., Davis, R. W., *Proc. Natl. Acad. Sci. USA* 1997, 94, 13057-13062.
- [20] Shalon, D., Smith, S. J., Brown, P. O., *Genome Res.* 1996, 6, 639-645.
- [21] Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., *Science* 1995, 270, 484-487.
- [22] Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., Kinzler, K. W., *Cell* 1997, 88, 243-251.
- [23] Krishna, R. G., Wold, P., *Adv. Enzymol.* 1993, 67, 265-298.
- [24] G6rg, A., Postel, W., Gunther, S., *Electrophoresis* 1988, 9, 531-546.
- [25] Klose, J., Kobalz, U., *Electrophoresis* 1995, 16, 1034-1059.
- [26] Matsudaira, P., *J. Biol. Chem.* 1987, 262, 10035-10038.
- [27] Aebersold, R. H., Teplow, D. B., Hood, L. E., Kent, S. B., *J. Biol. Chem.* 1986, 261, 4229-4238.
- [28] Rosenfeld, J., Capdevielle, J., Guillemot, J. C., Ferrara, P., *Anal. Biochem.* 1992, 203, 173-179.
- [29] Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., Kent, S. B., *Proc. Natl. Acad. Sci. USA* 1987, 84, 6970-6974.
- [30] Honor6, B., Leffers, H., Madsen, P., Celis, J. E., *Eur. J. Biochem.* 1993, 218, 421-430.
- [31] Mann, M., Wilm, M., *Anal. Chem.* 1994, 66, 4390-4399.
- [32] Eng, J., McCormack, A. L., Yates III, J. R., *J. Amer. Mass Spectrom.* 1994, 5, 976-989.
- [33] Yates III, J. R., Eng, J. K., McCormack, A. L., Schlett, D., *Anal. Chem.* 1995, 67, 1426-1436.
- [34] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, 68, 850-858.
- [35] Hess, D., Covey, T. C., Winz, R., Brownsey, R. W., Aebersold, R., *Protein Sci.* 1993, 2, 1342-1351.
- [36] van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Biochem.* 1997, 247, 310-318.
- [37] Lui, M., Tempst, P., Erdjument-Bromage, H., *Anal. Biochem.* 1996, 241, 156-166.
- [38] Patterson, S. D., Aebersold, R. A., *Electrophoresis* 1995, 16, 1791-1814.
- [39] Ducret, A., Foyn, Brunn, C., Bures, E. J., Marhaug, G., Husby, G. R. A., *Electrophoresis* 1996, 17, 866-876.
- [40] Haynes, P. A., Fripp, N., Aebersold, R., *Electrophoresis* 1998, 19, 939-945.
- [41] Figeys, D., Van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Chem.* 1996, 68, 1822-1828.
- [42] Ducret, A., Van Oostveen, I., Eng, J. K., Yates III, J. R., Aebersold, R., *Protein Sci.* 1997, 7, 706-719.
- [43] Figeys, D., Ducret, A., Yates III, J. R., Aebersold, R., *Nature Biotech.* 1996, 14, 1579-1583.
- [44] Figeys, D., Aebersold, R., *Electrophoresis* 1997, 18, 360-368.
- [45] Figeys, D., Ning, Y., Aebersold, R., *Anal. Chem.* 1997, 69, 3153-3160.
- [46] Garrels, J. I., McLaughlin, C. S., Warner, J. R., Fletcher, B., Latter, G. I., Kobayashi, R., Schwender, B., Volpe, T., Anderson, D. S., Mesquita-Fuentes, R., Payne, W. B., *Electrophoresis* 1997, 18, 1347-1360.
- [47] Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B. B., Butler, A., Castle, A. B., Chianlikulchai, N., Chu, A., Cleo, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., Edwards, C., Fan, J.-B., Fang, N., Fizames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hui, L., Hussain, S., Louis-Dit-Sully, C., Ma, J., MacGillivray, A., Mader, C., Maratukulam, A., Matise, T. C., McKusick, K. B., Morrisette, J., Mungall, A., Muselet, D., Nusbaum, H. C., Page, D. C., Peck, A., Perkins, S., Piercy, M., Qin, P., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, X., She, X., Silva, J., Slonim, D. K., Soderlund, C., Sun, W.-L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wilmer, T., Wu, X., Adams, M. D., Auffray, C., Walter, N. A. R., Brandon, R., Dehaja, A., Goodfellow, P. N., Houlgatte, R., Hudson, J. R., Jr., Ido, S. E., Iorio, K. R., Lee, W. Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M. H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J. C., Sikela, J. M., Beckmann, J. S., Weissenbach, J., Myers, R. M., Cox, D. R., James, M. R., Bentley, D., et al. *Science* 1996, 274, 540-546.
- [48] Sanchez, J.-C., Rouge, V., Pisteur, M., Ravier, F., Tonella, L., Moosmayer, M., Wilkins, M. R., Hochstrasser, D. F., *Electrophoresis* 1997, 18, 324-327.
- [49] Miller, I., Haynes, P., Gemeiner, M., Aebersold, R., Manzoni, C., Lovati, M. R., Vignati, M., Eberlin, I., Gianazza, E., *Electrophoresis* 1998, 19, 1493-1500.
- [50] Garrels, J. I., *Nucleic Acids Res.* 1996, 24, 46-49.



RECEIVED

NOV 24 2004

TECH CENTER 1600/2900

editorial

Journal of  
**proteome**  
research**EDITOR-IN-CHIEF****William S. Hancock**Barnett Institute and  
Department of Chemistry  
Northeastern University  
360 Huntington Avenue  
341 Mugar Bldg.  
Boston, MA 02115  
617-373-4881; Fax: 617-373-2855  
whancock@acs.org**ASSOCIATE EDITORS****Joshua LaBaer**

Harvard Medical School

**György Marko-Varga**

AstraZeneca and Lund University

**EDITORIAL ADVISORY BOARD****Ruedi H. Aebersold**

Institute for Systems Biology

**Leigh Anderson**

Plasma Proteome Institute

**Ettore Appella**

National Cancer Institute

**Rolf Apweiler**

European Bioinformatics Institute

**Ronald Beavis**

Manitoba Centre for Proteomics

**Walter Blackstock**

Cellzome

**Brian Chait**

The Rockefeller University

**Patrick L. Coleman**

3M

**Christine Colvis**

National Institutes of Health

**Catherine Fenselau**

University of Maryland

**Daniel Figeys**

MDS Proteomics

**Sam Hanash**

University of Michigan

**Stanley Hefta**

Bristol-Myers Squibb

**Donald F. Hunt**

University of Virginia

**Barry L. Karger**

Northeastern University

**Daniel C. Liebler**

Vanderbilt University School of Medicine

**Lance Liotta**

National Cancer Institute

**Matthias Mann**

University of Southern Denmark

**Stephen A. Martin**

Applied Biosystems

**Jeremy Nicholson**

Imperial College of London

**Gilbert S. Omenn**

University of Michigan

**Emanuel Petricoin**

Food and Drug Administration

**J. Michael Ramsey**

Oak Ridge National Laboratory

**Pier Giorgio Righetti**

University of Verona

**John T. Stults**

Biospect

**Peter Wagner**

Zyomix

**Keith Williams**

Proteome Systems

**Qi-Chang Xia**

Shanghai Institute of Biochemistry

**John R. Yates, III**

The Scripps Research Institute

*Do We Have Enough Biomarkers?*

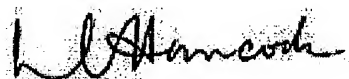
**T**he Editor has become aware of a recent push to validate currently available biomarkers in an extensive clinical setting. The reasoning behind such a push is that there are already a significant number of biomarkers that now need to be used effectively in the clinic. Many biomarkers, such as the carcinoembryonic antigen, have been known for some time and are used widely for patient management. The older biomarkers, however, are not effective for early diagnosis.

With the advent of genomics and, later, proteomics, there has been a substantial investment in using these new tools to generate additional biomarkers. The problem with this new information is that it is too early to get consensus on what is a useful marker or what is a good patient population for such a study. Therefore, it is unclear whether the new markers currently in hand will give better clinical information than the ones that have been used in the past. An additional problem is that the markers that are generated by proteomics are not always consistent with the markers that are generated from expression profiling.

The challenge in this situation is to balance the need of patients for better, early diagnosis of disease with the need to have high-quality markers for the expensive and time-consuming validation process. This Editor believes that proteomics is at too early a stage for this new technology to have generated a quality list of markers. The risk is if we push the existing markers into extensive clinical validation, we will be missing the fruits of improvements in emerging proteomics technology. I think many people in the proteomics community would agree that federal granting agencies should be enticed to continue investments in basic proteomics technology. In addition, funding should be made available for basic science studies that will continue to generate biomarkers, and there needs to be some type of consensus-building process that can lead to a consolidation of the different lists of biomarkers.

There are good past models for such activities, such as the consensus-forming meetings that the U.S. Food and Drug Administration has held; these yielded technical innovations. One example was the generation of new protein pharmaceuticals at the advent of the biotechnology industry. Another example, in the early days of the genome sequencing program, was when a group of experts came together to agree on annotation of the early results. The Human Proteome Organization is a good example of an international group of laboratories coming together to consolidate the output from a number of studies with different technology platforms.

I would like to encourage the biomedical community not to rush to judgment in terms of biomarkers, but instead to give research more time to produce quality biomarker information. Then we should conduct a thorough evaluation of a widely agreed-on list before we attempt to determine which of these new markers are indeed worthy of extensive clinical investigation.



## Analysis of Genomic and Proteomic Data Using Advanced Literature Mining

Yanhui Hu, Lisa M. Hines, Haifeng Weng, Dongmei Zuo, Miguel Rivera,  
Andrea Richardson, and Joshua LaBaer\*

*Institute of Proteomics, Harvard Medical School-BCMP, 240 Longwood Avenue, Boston, Massachusetts 02115*

Received March 13, 2003

High-throughput technologies, such as proteomic screening and DNA micro-arrays, produce vast amounts of data requiring comprehensive analytical methods to decipher the biologically relevant results. One approach would be to manually search the biomedical literature; however, this would be an arduous task. We developed an automated literature-mining tool, termed MedGene, which comprehensively summarizes and estimates the relative strengths of all human gene-disease relationships in Medline. Using MedGene, we analyzed a novel micro-array expression dataset comparing breast cancer and normal breast tissue in the context of existing knowledge. We found no correlation between the strength of the literature association and the magnitude of the difference in expression level when considering changes as high as 5-fold; however, a significant correlation was observed ( $r = 0.41$ ;  $p = 0.05$ ) among genes showing an expression difference of 10-fold or more. Interestingly, this only held true for estrogen receptor (ER) positive tumors, not ER negative. MedGene identified a set of relatively understudied, yet highly expressed genes in ER negative tumors worthy of further examination.

**Keywords:** bioinformatics • micro-array • text mining • gene-disease association • breast cancer

### Introduction

At its current pace, the accumulation of biomedical literature outpaces the ability of most researchers and clinicians to stay abreast of their own immediate fields, let alone cover a broader range of topics. For example, to follow a single disease, e.g., breast cancer, a researcher would have had to scan 130 different journals and read 27 papers per day in 1999.<sup>1</sup> This problem is accentuated with high-throughput technologies such as DNA micro-arrays and proteomics, which require the analysis of large datasets involving thousands of genes, many of which are unfamiliar to a particular researcher. In any microarray experiment, thousands of genes may demonstrate statistically significant expression changes, but only a fraction of these may be relevant to the study. The ability to interpret these datasets would be enhanced if they could be compared to a comprehensive summary of what is known about all genes. Thus, there is a need to summarize existing knowledge in a format that allows for the rapid analysis of associations between genes and diseases or other specific biological concepts.

One solution to this problem is to compile structured digital resources, such as the Breast Cancer Gene Database<sup>1</sup> and the Tumor Gene Database.<sup>2</sup> However, as these resources are hand-curated, the labor-intensive review process becomes a rate-limiting step in the growth of the database. As a result, these

databases have a limited scale and the genes are not selected in a systematic fashion.

An alternative approach is automated text mining; a method which involves automated information extraction by searching documents for text strings and analyzing their frequency and context. This approach has been used successfully in several instances for biological applications. In most cases, it has been applied to extract information about the relationships or interactions that proteins or genes have with one another, in the literature or by functional annotation.<sup>3-7</sup> Thus far, few publications have applied text-mining to examine the global relationships between genes and diseases. Perez-Iratxeta et al. automatically examined the GO (Gene Ontology) annotation of genes and their predicted chromosomal locations in order to identify genes linked to inherited disorders.<sup>8</sup>

To obtain a more global understanding of disease development, it would be valuable to incorporate information regarding all possible gene-disease relationships, including biochemical, physiological, pharmacological, epidemiological, as well as genetic. This information would enable comprehensive comparisons between large experimental datasets and existing knowledge in the literature. This would accomplish two things. First, it would serve to validate experiments by demonstrating that known responses occur as predicted. Second, it would rapidly highlight which genes are corroborated by the literature and which genes are novel in a given context. We have utilized a computational approach to literature mining to produce a

\* To whom correspondence should be addressed: jlabar@hms.harvard.edu.

comprehensive set of gene-disease relationships. In addition, we have developed a novel approach to assess the strength of each association based on the frequency of citation and co-citation. We applied this tool to help interpret the data from a large micro-array gene expression experiment comparing normal and cancerous breast tissue.

## Methods

**MedGene Database.** MedGene is a relational database, storing disease and gene information from NCBI, text mining results, statistical scores, and hyperlinks to the primary literature. MedGene has a web-based user interface for users to query the database (<http://hipseq.med.harvard.edu/MedGene/>).

**Text Mining Algorithms.** MeSH files were downloaded from the MeSH web site at NLM (National Library of Medicine) (<http://www.nlm.nih.gov/mesh/meshhome.html>) and human disease categories were selected. LocusLink files were downloaded from the LocusLink web site at NCBI (<http://www.ncbi.nih.gov/LocusLink/>). Official/preferred gene symbol, official/preferred gene name, and gene alternative symbols and names, all relevant annotations and URLs for each LocusLink record, were collected. Gene search terms were used for literature searching and included all qualified gene names, gene symbols, and gene family terms. Primary gene keys, predominantly qualified gene family terms and gene official/preferred symbols, were used to index Medline records. If the official/preferred gene symbols did not meet the standards to be an index, then qualified gene official/preferred names were used. A local copy of Medline records (up to July, 2002) was pre-selected.

A JAVA module examined the MeSH terms and then indexed each Medline record with the appropriate disease terms. A separate JAVA module was used to examine the titles and abstracts for gene search terms and then to index the gene-related Medline records with the relevant primary gene key(s).

**Statistical Methods.** For every gene and disease pair, we counted records that were indexed for both gene and disease (double positive hits), for disease only (disease single hits), for gene only (gene single hits), and for neither gene nor disease (double negative hits) to generate a  $2 \times 2$  contingency table. On the basis of the contingency table-framework, we applied different statistical methods to estimate the strength of gene-disease relationships and evaluated the results. These methods included chi-square analysis, Fisher's exact probabilities, relative risk of gene, and relative risk of disease<sup>16</sup> (<http://hipseq.med.harvard.edu/MedGene/>). In addition, we computed the "product of frequency", which is the product of the proportion of disease/gene double hits to disease single hits and the proportion of disease/gene double hits to gene single hits. To obtain a normal distribution, we transformed all the statistical scores using the natural logarithm. We selected the log of the product of frequency (LPF) to validate MedGene and to use for the analysis with the micro-array data. Spearman rank-correlation coefficients were used to assess the linear relationship between LPF and micro-array fold change in expression level.

**Global Analysis.** Diseases with at least 50 related genes were selected for clustering analysis, and the LPF scores were normalized with total score for each disease. Hierarchical clustering was done with the "Cluster" software and the clustering result was visualized using "TreeView" (<http://rana.lbl.gov/EisenSoftware.htm>).

**Breast Tissue Micro-Arrays.** Eighty-nine breast cancer samples (79% ER-positive) and 7 normal breast tissue samples were selected from the Harvard Breast SPORE frozen tissue repository and were representative of the spectrum of histological types, grades, and hormone receptor immuno-phenotypes of breast cancer. Biotinylated cRNA, generated from the total RNA extracted from the bulk tumor, was hybridized to Affymetrix U95A oligo-nucleotide micro-arrays. These micro-arrays consist of 12 400 probes, which represent approximately 9000 genes. Raw expression values were obtained using GENE-CHIP software from Affymetrix, and then further analyzed using the DNA-Chip Analyzer (dChip) custom software.

## Results

**Automated Indexing of Medline Records by Disease and Gene.** To study the gene-disease associations in the literature, we first compiled complete lists for human diseases and human genes. To index all Medline records that were relevant to human diseases, the Medical Subject Heading (MeSH) index of Medline records was utilized. MeSH is a controlled medical vocabulary from the National Library of Medicine and consists of a set of terms or subject headings that are arranged in both an alphabetic and an hierarchical structure. Medline records are reviewed manually and MeSH terms are added to each with software assistance.<sup>9,10</sup> Twenty-three human disease category headings along with all of their child terms (see the Supporting Information, Supplemental Table 1, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Table1.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Table1.html)) were selected from the 2002 MeSH index creating a list of 4033 human diseases.

No index comparable to the MeSH index exists for genes, and thus, it was necessary to apply a string search algorithm for gene names or symbols found in Medline text. A complete list of genes, gene names, gene symbols, and frequently used synonyms were collected from the LocusLink database at NCBI,<sup>11,12</sup> which contains 53 259 independent records keyed by an official gene symbol or name (June 18<sup>th</sup>, 2002). For the purposes of this study, no distinction was made between genes and their gene products. Authors often use the same name for both, differentiating the two only by the use of italics, if at all. For the intended use of this study, this lack of distinction is unlikely to have a large effect and may in fact be beneficial.

Initial attempts to search the literature using these lists revealed several sources of false positives and false negatives (Table 1). False positives primarily arose when the searched term had other meanings, whereas false negatives arose from syntax discrepancies necessitating the development of filters to reduce these errors. The syntax issues were readily handled by including alternate syntax forms in the search terms. The false positive cases, caused by duplicative and unrelated meanings for the terms, were more difficult to manage. Where possible, case sensitive string mapping reduced inappropriate citations. In many cases, however, this was not sufficient and the terms had to be eliminated entirely, thereby reducing the false positive rate but unavoidably under-representing some genes.

For the purposes of data tracking, a primary gene key was selected to represent all synonyms that correspond to each gene. Medline records were indexed with a primary gene key when any synonym for that key was found in the title or abstract. Case-insensitive string mapping was used for all searches except as noted above. No additional weight was

**Table 1.** Systematic Sources of False Positives and False Negatives in Unfiltered Data<sup>a</sup>

source of error	error type	example	filter solution
gene symbol/name is not unique	false positive	<i>MAG</i> —myelin associated glycoprotein <i>MAG</i> —malignancy-associated protein	eliminate this term
gene symbol is unrelated abbreviation	false positive	<i>PA</i> —pallid homologue (mouse), pallidin (also abbrev. for Pennsylvania)	eliminate this term
gene symbol/name has language meaning	false positive	<i>WAS</i> —Wiskott–Aldrich Syndrome (also the word “was”)	case-sensitive string search
nonstandard syntax	false negative	<i>BAG-1</i> instead of <i>BAG1</i>	add dash term
unofficial gene name/symbol	false negative	<i>P53</i> instead of <i>TP53</i>	add all gene nicknames
nonspecified gene name	false negative	estrogen receptor instead of Estrogen receptor 1	add family stem term

<sup>a</sup> In preliminary studies, Medline was searched for co-occurrence of genes and diseases and the resulting output was evaluated to identify error sources that were amenable to global filters. Each error source is categorized by the type of error it causes: false positives are suggested relationships that are not real and false negatives are real relationships that are underrepresented. The filter solutions used are indicated. Note that in some cases, the filter solution itself introduces error. In general, error rates maximized sensitivity, even at the expense of specificity if needed.

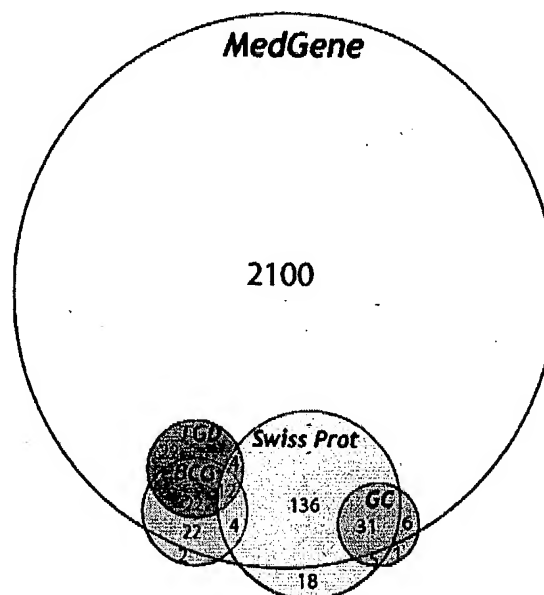
added for multiple occurrences of a term or the co-occurrence of multiple synonyms for the same gene key.

Medline records were searched with all qualified gene identifiers, such as the official/preferred gene symbol, the official/preferred gene name, all gene nicknames and all syntax variants. In situations where there are several members of a gene family or splice variants, some authors prefer to use a shortened gene family name, e.g., estrogen receptor instead of estrogen receptor 1 (*ESR1*), creating a source of false negatives. For this reason, gene family stem terms were created for all genes that have an alpha or numerical suffix (e.g., *IL2RA*, *TGFβ*, *ESR1*, etc.) and then used to search the literature. The family stem terms were handled separately from the specific gene names so that it would be clear when linkages were made to the gene family versus a specific member in that family.

To improve performance and accuracy, some pre-selection was applied to the records that were scanned. First, review articles were eliminated to avoid redundant treatment of citations. Second, non-English journals were removed because the natural language filters were only relevant to English publications. Finally, journals unlikely to contain primary data about gene-disease relationships were also removed (e.g., *Int. J. Health Educ.*, *Bedside Nurse*, and *J. Health Econ.*). Together, these filters reduced the 12 198 221 Medline publications (July 2002) by 37%.

**Ranking the Relative Strengths of Gene-Disease Associations.** In total, there were 618 708 gene-disease co-citations, in which 16% (8297) of all studied genes had been associated to a disease and 96% (3875) of all diseases had been associated to at least one gene. To rank the relative strengths of gene disease relationships, we tested several different statistical methods and examined the results. With the exception of the relative risk estimates, the methods provided similar results with respect to the rank order of the gene-disease association strengths. However, after comparing the results to other databases and after consulting disease experts, the log of the product of frequency (LPF) was selected for further analysis because it gave the best results overall.

**Validation of MedGene.** In developing this tool, it was important to minimize the number of missed genes (false negatives) and miscalled genes (false positives). However, in situations when these goals were in conflict, inclusiveness was prioritized. To determine the false negative rate in MedGene, breast cancer was used as a test case because it was associated with more genes than any other human disease and because



**Figure 1.** Estimation of the false negative rate by comparison with hand-curated databases. The breast cancer-related genes identified by MedGene were compared with those listed in several other databases including the Tumor Gene Database (TGD),<sup>2</sup> the Breast Cancer Gene Database (BCG),<sup>1</sup> GeneCards (GC)<sup>17</sup> and Swissprot.<sup>18</sup> Genes were considered false negatives if they were represented in at least one of these other databases and not in MedGene and their link to breast cancer was supported by at least one literature reference. All literature references were verified by manual review to confirm their validity. The number of genes in each database or shared by more than one database is indicated. The false negative rate was calculated by genes missed at MedGene (26)/total number of nonoverlapping genes in other databases (285).

there were several public databases that link genes to breast cancer. We compared the list of breast cancer-related genes from MedGene to these databases, illustrated in Figure 1. Among the 285 distinct breast cancer-related genes that were supported by at least one literature citation in these hand-curated databases, 26 were absent from MedGene, suggesting a false negative rate of approximately 9%. To determine why these were missed, all literature references for these genes (80

papers) were reviewed manually (see the Supporting Information, Supplemental Table 2, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Table 2.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Table 2.html)). Among these papers, most false negatives were caused by nonstandard gene terms or gene terms eliminated by our specificity filters. Few genes were missed because they were only mentioned in review papers (0.4%) or they appeared only in the body of the manuscript but not the abstract or title (1.1%). Of note, MedGene identified approximately 2000 additional breast cancer-related genes not listed in any other database.

To assess the false positive error rate, two complementary approaches were used: a detailed analysis of one disease and a global examination of 1000 diseases. The detailed approach examined the false positive error rate and its sources, whereas the global approach tested whether the overall results made biomedical sense.

Using the LPF, 1467 genes related to prostate cancer were assembled in rank order. We then retrieved approximately 300 Medline records each for the highest ranked 100 and the lowest ranked 200 genes and manually reviewed the titles and abstracts to determine the verity of the association. Nearly 80% of the highest ranked 100 genes fell into one of the five categories that reflect meaningful gene-disease relationships (see the Supporting Information, Supplemental Table 3, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Table 3.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Table 3.html)). Among the lowest ranked 200 genes, approximately 70% reflected true relationships. Of the 600 records reviewed, there were only two in which the association between the gene and the disease was described as negative. Both were genes with very low scores. In both cases, the authors did not argue the absence of any relationship, but rather that a particular feature of the gene or protein was not shown to be related to human prostate cancer.<sup>13,14</sup>

The coincidence of some gene symbols with medical abbreviations, chemical abbreviations and biological abbreviations resulted in most of the false positives (see the Supporting Information, Supplemental Table 4, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Table 4.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Table 4.html)), emphasizing the importance of the filters that were added in the search algorithm (Table 1). Without the filters, the false positive rate more than doubled, and the false negative rate rose dramatically (data not shown). For example, among the papers about breast cancer, there were only 12 Medline records that referred to *ESR1* and 10 to *ESR2*, whereas almost 2000 papers mentioned estrogen receptor without specifying *ESR1* or *ESR2*; this latter group was detected by the family stem term filter.

To further validate these results, a global analysis of the gene-disease relationships described by MedGene was performed. For this experiment, it was reasoned that the more closely related the diseases are to one another, the more they will be related to the same gene sets. Thus, if the relationships defined by MedGene accurately reflected the literature, then an unsupervised hierarchical clustering of the gene data should group diseases in a manner consistent with common medical thinking. Conversely, if the clustered diseases do not make sense biologically or medically, it may reflect excessive false positives, false negatives, or inappropriate scoring of the data.

To execute this experiment, the gene sets and the corresponding LPF values for 1000 randomly selected diseases (each with at least 50 gene relationships) were used as a dataset for clustering the diseases. A review of the results showed that the resulting disease clusters were indeed logical based upon common medical knowledge (see the Supporting Information,

Supplemental Figure 1, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Figure 1.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 1.html)). For example, in one such cluster shown in Figure 2, diabetes and its complications grouped together and were also closely linked to diseases associated with starvation states.

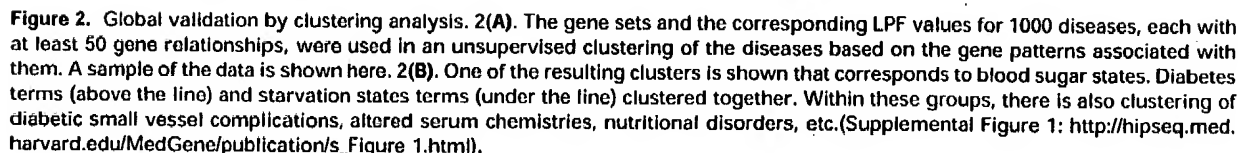
The number of genes associated with a given disease can be estimated by adjusting the MedGene number up by the false negative rate (~9%) and down by the false positive rate (~26% on average). Using this, the average disease has  $103.7 \pm 45.3$  (mean  $\pm$  s.d.) genes associated with it, although the range is quite broad with 2359 genes related to breast cancer, 2122 genes related to lung cancer and no genes related to a number of diseases.

**Applying MedGene to the Analysis of Large Datasets.** Access to a comprehensive summary of the genes linked to human diseases provided an opportunity to analyze data obtained from a high-throughput experiment. We compared the MedGene breast cancer gene list to a gene expression data set generated from a micro-array analysis comparing breast cancer and normal breast tissue samples. Micro-array analysis identified 2286 genes that had greater than a 1-fold difference in mean expression level between breast cancer samples and normal breast samples. Using MedGene, we sorted the 2286 genes into four classes: 555 genes directly linked to breast cancer in the literature by gene term search (first-degree association by gene name); 328 genes directly linked by family term search (first-degree association by family term); 1021 genes linked to breast cancer only through other breast cancer genes (second-degree association); and 505 genes not previously associated with breast cancer. (See the Supporting Information, Supplemental Figure 2, or visit [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Figure 2.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Figure 2.html).) Among the 505 previously unrelated genes, 467 were either newly identified genes or genes that had not previously been associated with any disease. Among the remaining 38 genes, 9 had been related to other cancers, specifically esophageal, colon, uterine, skin, and cervix.

To determine whether the genes highlighted by the micro-array analysis were more likely to have been previously linked to breast cancer in the literature, we created a two-dimensional plot of the fold change of expression level between breast cancer and normal tissue versus the literature score (LPF) (Figure 3A). There was a broad spread of expression changes among the genes directly linked to breast cancer ranging from less than 1-fold change (68%) to over 40-fold (0.3%). Notably, the majority of genes with greater than 10-fold expression changes were linked to breast cancer by first-degree association.

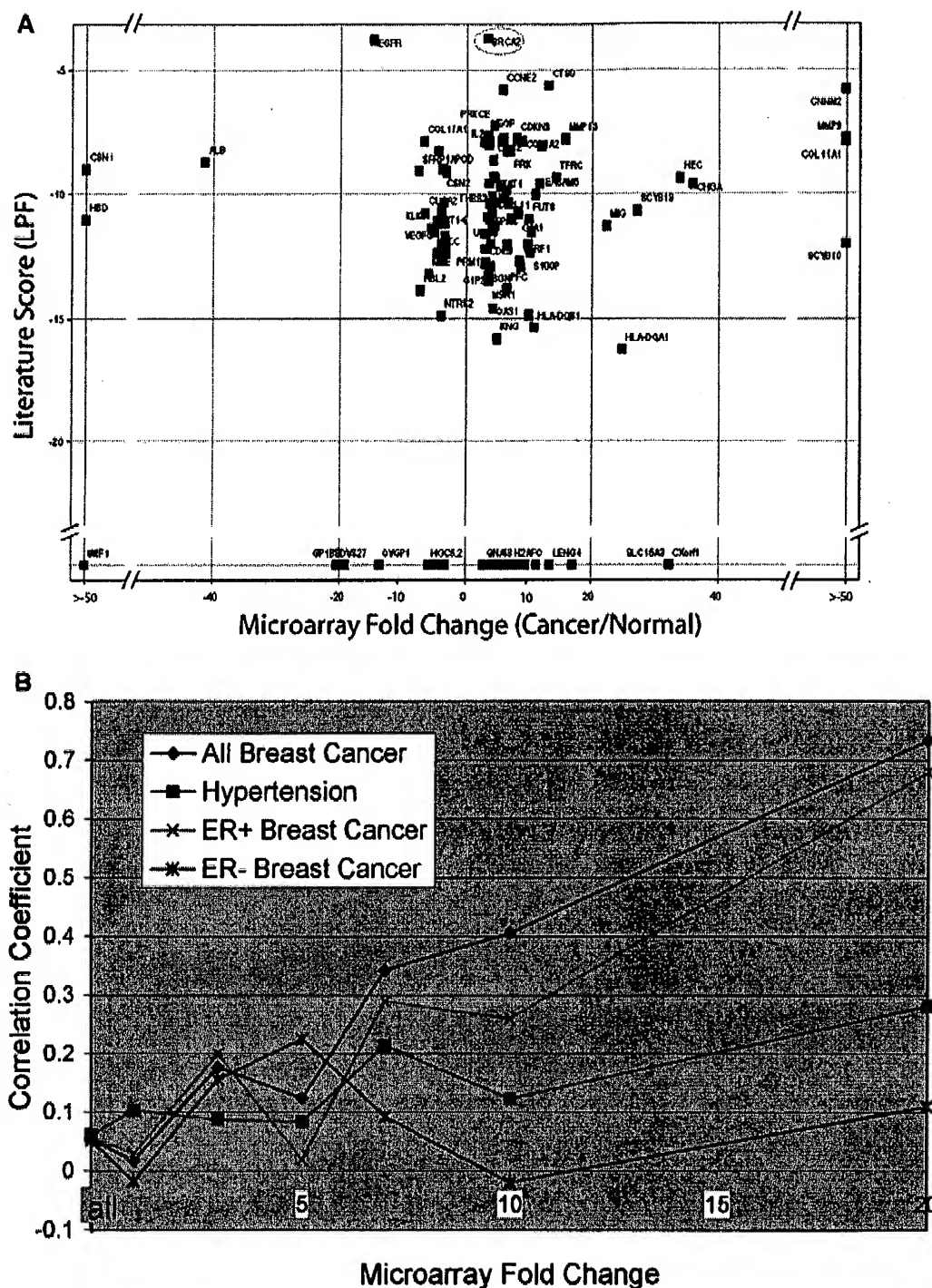
Among all 754 genes directly linked to breast cancer in the literature, there was no correlation between LPF and micro-array fold change ( $r = 0.018$ ,  $p$ -value = 0.62). However, when we stratified the analysis based on the magnitude of the fold change, we observed an increasing trend in correlation (Figure 3B) suggesting that genes with a more substantial change in expression level were more likely to have a stronger association in the literature. For genes that had 10-fold change or more in expression level, the correlation increased to 0.41 ( $p$ -value = 0.05).

When we evaluated the micro-array data separately for ER positive and ER negative tumors, the trend in correlation between fold change and literature score was highly dependent on estrogen receptor status. Interestingly, there was a similar trend in correlation for ER positive tumors, but no trend in correlation for ER negative tumors.



disease unrelated to breast cancer. As expected, we did not observe an increasing trend in correlation for hypertension.





**Figure 3.** Relationship between literature score and functional data for breast cancer. **3A.** The data from an expression analysis of samples for breast tumors and normal breast tissue were analyzed to indicate the fold difference of expression level between breast tumor and normal sample (cutoff  $\geq 3$ -fold change). The fold changes were plotted against the literature score for the same gene set. Green dots represent first-degree association by gene search, blue dots represent first-degree association by family search and red dots represent no-association. Some well-studied genes, such as BRCA2 (pink circle), are not reflected by a substantial difference in expression level. Furthermore, the majority of genes that have no association with breast cancer in the literature had less than 10-fold expression changes (shaded area). **3B.** The Spearman rank-correlation coefficients between literature score (LPF) and the fold change of expression level between tumor and normal breast samples (y-axis) in relation to the amount of fold change of expression level (x-axis). Gene rank lists were generated for breast cancer (blue) and hypertension (pink). Correlations were also computed between the breast cancer gene LPF scores and fold change expression data among estrogen receptor positive tumors only (light blue) and estrogen receptor negative tumors only (purple).

Table 2. Top 25 Genes Related to Selected Human Diseases<sup>a</sup>

breast neoplasms	hypertension	rheumatoid arthritis	bipolar disorder	atherosclerosis
estrogen receptor	<i>REN</i>	<i>RA</i>	<i>ERDA1</i>	apolipoprotein
<i>PGR</i>	<i>DBP</i>	<i>TNFRSF10A</i>	<i>SNAP29</i>	<i>APOE</i>
<i>ERBB2</i>	<i>LEP</i>	<i>CRP</i>	<i>PFKL</i>	<i>LDLR</i>
<i>BRCA1</i>	<i>AGT</i>	<i>AS</i>	<i>DRD2</i>	<i>ELN</i>
<i>BRCA2</i>	<i>INS</i>	<i>ESR1</i>	<i>TRH</i>	<i>ARG1</i>
<i>EGFR</i>	kallikrein	<i>HLA-DRB1</i>	<i>IMPA2</i>	<i>APOB</i>
<i>CYP19</i>	<i>ACE</i>	<i>DR1</i>	<i>HTR3A</i>	<i>APOA1</i>
<i>TFF1</i>	endothelin	interleukin	<i>DRD3</i>	<i>MSR1</i>
<i>PSEN2</i>	<i>S100A6</i>	<i>TNF</i>	<i>REM</i>	<i>LPL</i>
<i>TP53</i>	<i>BDK</i>	<i>IL6</i>	<i>KCNN3</i>	<i>PON1</i>
<i>CES3</i>	<i>DIANPH</i>	collagen	<i>DRD4</i>	plasminogen
<i>CEACAM5</i>	<i>SAR1</i>	<i>IL1A</i>	<i>HTR2C</i>	activator inhibitor
<i>ERBB3</i>	<i>PIH</i>	<i>ACR</i>	<i>RELN</i>	<i>PLG</i>
cyclin	<i>CD59</i>	<i>TNFRSF12</i>	<i>DBH</i>	vascular cell
<i>COX5A</i>	<i>ALB</i>	<i>IL2</i>	<i>MAOA</i>	adhesion molecule
cathepsin	<i>CYP11B2</i>	<i>CHI3L1</i>	<i>COMT</i>	<i>ATOH1</i>
<i>ERBB4</i>	<i>MAT2B</i>	<i>IL8</i>	<i>HTR2A</i>	<i>VWF</i>
<i>TRAM</i>	angiotensin receptor	interleukin 1 matrix	<i>SYNJ1</i>	<i>INS</i>
<i>CCND1</i>	<i>AGTR2</i>	metalloproteinase	<i>INPP1</i>	<i>ARG2</i>
<i>EGF</i>	<i>NPPA</i>	interferon	<i>NEDD4L</i>	<i>ABCA1</i>
<i>MUC1</i>	<i>LVM</i>	<i>CD68</i>	<i>FRA13C</i>	<i>OLR1</i>
insulin-like	<i>DBH</i>	<i>IL4</i>	transducer of	collagen
<i>BCL2</i>	<i>NPY</i>	<i>IL17</i>	<i>ERBB2</i>	<i>MCP</i>
mucin	<i>POMC</i>	<i>MMP3</i>	<i>BAIAP3</i>	lipoprotein
<i>FGF3</i>	neuropeptide	<i>SIL</i>	<i>ATP1B3</i>	<i>APOA2</i>
			<i>DRD5</i>	intercellular
				adhesion molecule
				<i>RAB27A</i>

<sup>a</sup> MedGene results for the top 25 genes associated with breast neoplasms, hypertension, rheumatoid arthritis, bipolar disorder, and atherosclerosis, respectively, ranked by LPF scores. The hyperlink to all the papers co-citing the gene and the disease is available at MedGene website (<http://hipseq.med.harvard.edu/MedGene/>).

## Discussion

The Human Genome Project heralded a new era in biological research where the emphasis on understanding specific pathways has expanded to global studies of genomic organization and biological systems. High-throughput technologies can provide novel insight into comprehensive biological function but also introduces new challenges. The utility of these technologies is limited to the ability to generate, analyze, and interpret large gene lists. MedGene, a relational database derived by mining the information in Medline, was created to address this need. MedGene users can query for a rank-ordered list of human gene-disease relationships (Table 2) for one or more diseases. Each entry is hyperlinked to the original papers supporting each association and to other relevant databases.

MedGene is an innovative extension of previous text mining approaches. Perez-Iratxeta et al. used the GO annotation and their chromosomal locations to predict genes that may contribute to inherited disorders.<sup>8</sup> MedGene takes a broader view and includes all diseases and all possible gene-disease relationships. Furthermore, MedGene utilizes co-citation to indicate a relationship rather than GO annotation, which is limited to the subset of genes that have GO annotation. Our approach is complementary to that taken by Chaussabel and Sher, who used the frequency of co-cited terms to cluster genes into a hierarchy of gene-gene relationships.<sup>6</sup>

A unique aspect of this tool is the ability to assess the relative strengths of gene-disease relationships based on the frequency of both co-citation and single citation. This presupposes that most co-citations describe a positive association, often referred to as publication bias<sup>15</sup> and is supported by our observations

that negative associations are rare (Supplemental Table 3: [http://hipseq.med.harvard.edu/MedGene/publication/s\\_Table3.html](http://hipseq.med.harvard.edu/MedGene/publication/s_Table3.html)). Of course, relationships established by frequency of co-citation do not necessarily represent a true biological link; however, it is strong evidence to support a true relationship.

Another important feature of MedGene is the implementation of software filters that substantially reduced the error rate. We estimate that less than 10% of all associations were missed and at least 70% of even the weakest associations were real. For this study, all of the filters that we applied were general ones, e.g., expanding the list of all gene names to address the different syntax forms used by different journals, eliminating gene names that correspond to common English words, etc. The majority of the remaining search term ambiguities were idiosyncratic and difficult to identify systematically without causing a significant rise in false negatives. Alternative approaches, such as the examination of the nearest neighbor terms, need to be considered to further reduce the false positive rate.

It is not uncommon to see expression changes in microarray experiments as small as 2-fold reported in the literature. Even when these expression changes are statistically significant, it is not always clear if they are biologically meaningful. When comparing expression levels of disease to normal tissue, one expects an enrichment of known disease-related genes to appear in the altered expression group. MedGene provided a unique opportunity to test this notion in the context of existing knowledge on a novel breast cancer micro-array dataset. For genes displaying a 5-fold change or less in tumors compared to normal, there was no evidence of a correlation between altered gene expression and a known role in the disease. This



**Table 3.** Genes with Large Expression Changes in ER- but Not in ER+ Breast Tumors

gene symbol	fold change (ER+)	fold change (ER-)
KRTHB1	1.0	610.8
BRS3	1.2	89.4
DKK1	1.2	69.8
ZIC1	1.9	59.6
TLR1	1.0	38.5
KIAA0680	2.6	33.2
CDKN3	1.0	30.6
EBI2	4.0	27.9
GZMB	3.8	21.9
STK18	4.7	18.6
GPR49	1.0	14.6
MYO10	1.6	14.4
LAD1	-1.0	13.5
POLE2	4.2	13.0
HMG4	4.4	12.9
BCL2L11	-1.2	12.3
LRP8	2.9	12.2
CCNB2	1.0	11.8
CCNE2	4.0	11.6
FGF	-4.3	11.1
KNSL6	2.9	10.9
HIF5	3.0	10.2
SERPINH2	4.6	10.2
YAP1	1.0	10.0
LPHB	-1.3	-10.4
TCEA2	-1.1	-10.8
TFF1	1.3	-11.4
COL17A1	-4.1	-15.7
POP5	1.1	-16.2
BPAG1	-4.6	-22.3
PDZK1	-1.1	-36.8
VEGFC	-2.8	-51.5
MUC6	-1.4	-64.9
SERPINA5	-1.0	-83.1
MEIS1	-1.6	-85.9
CA12	2.4	-150.3

Table 3. MedGene identified a set of relatively understudied, yet highly expressed genes in ER negative, but not ER positive breast tumors. All of these genes have either never been co-cited with breast cancer or have a weak association except those marked with an \*.

reflects the many genes whose role in breast cancer may not involve large changes in expression in sporadic tumors (e.g., *BRCA1* and *BRCA2*) and genes whose modest changes in expression may be unrelated to the disease. Strikingly, among genes with a 10-fold change or more in expression level, there was a strong and significant correlation between expression level and a published role in the disease, providing the first global validation of the micro-array approach to identifying disease-specific genes.

The results derived from MedGene have two implications. First, a careful hunt for corroborating evidence of a role in breast cancer should precede any further study of genes with less than 5-fold expression level changes. Second, any genes with 10-fold changes or more are likely to be related to breast cancer and warrant attention. It is likely that this threshold will change depending on the disease as well as the experiment.

Interestingly, the observed correlation was only found among ER-positive tumors, not ER-negative. This may reflect a bias in the literature to study the more prevalent type of tumor in the population. Furthermore, this emphasizes that caution must be taken when interpreting experiments that may contain subpopulations that behave very differently. The MedGene approach identified a set of relatively understudied, yet highly expressed genes in ER-negative tumors that are worthy of further examination (Table 3).

In conclusion, we have developed an automated method of summarizing and organizing the vast biomedical literature. To our knowledge, the resulting database is the most comprehensive and accurate of its kind. By generating a score that reflects the strength of the association, it provides an important tool for the rapid and flexible analysis of large datasets from various high-throughput screening experiments. Furthermore, it can be used for selecting subsets of genes for functional studies, for building disease-specific arrays, for looking at genes common to multiple diseases and various other high-throughput applications. In the future, it will be possible to enhance the utility of the MedGene database by building links between genes and other MeSH terms as well as other biological processes and concepts, such as cell division and responses to small molecules.

**Acknowledgment.** We would like to thank P. Braun, L. Garraway, J. Pearlberg, and other members of our Institute for helpful discussion. Many thanks to the NLM (National Library of Medicine) for licensing of MEDLINE and the annotation effort of adding MeSH indexes for MEDLINE abstracts. This work was funded by grants from the Breast Cancer Research Foundation and an NHLBI PGA Grant (Vol HL66582-02).

**Supporting Information Available:** Twenty-three human disease category headings along with all of their child terms selected from the 2002 MeSH index (Supplemental Table 1); analysis of the causes of false negatives in MedGene (Supplemental Table 2); meaningful gene-disease relationships found in MedGene (Supplemental Table 3); causes for incorrect assignment of gene indexes (Supplemental Table 4); a review of the results, showing that the resulting disease clusters were indeed logical (Supplemental Figure 1); and a review of the results showing that among the 505 previously unrelated genes, 467 were either newly identified genes or genes that had not previously been associated with any disease (Supplemental Figure 2). This material is available free of charge via the Internet at <http://pubs.acs.org> and at the web sites mentioned in the text.

## References

- Basiri, R. A.; Glasser, S. R.; Steffen, D. L.; Wheeler, D. A. *Oncogene* **1999**, *18*, 7958-7965.
- Steffen, D. L.; Levine, A. E.; Yarus, S.; Basiri, R. A.; Wheeler, D. A. *Bioinformatics* **2000**, *16*, 639-649.
- Marcotte, E. M.; Xenarios, I.; Eisenberg, D. *Bioinformatics* **2001**, *17*, 359-363.
- Ono, T.; Hishigaki, H.; Tanigami, A.; Takagi, T. *Bioinformatics* **2001**, *17*, 155-161.
- Jenssen, T. K.; Laegreid, A.; Komorowski, J.; Hovig, E. *Nat. Genet.* **2001**, *28*, 21-28.
- Chaussabel, D.; Sher, A. *Genome Biol.* **2002**, *3*, RESEARCH0055.
- Gibbons, F. D.; Roth, F. P. *Genome Res.* **2002**, *12*, 1574-1581.
- Perez-Iratxeta, C.; Bork, P.; Andrade, M. A. *Nat. Genet.* **2002**, *31*, 316-319.
- Funk, M. E.; Reld, C. A. *Bull. Med. Libr. Assoc.* **1983**, *71*, 176-183.
- Humphrey, S. M.; Miller, N. E. *J. Am. Soc. Inf. Sci.* **1987**, *38*, 184-196.
- Maglott, D. R.; Katz, K. S.; Sicotte, H.; Pruitt, K. D. *Nucleic Acids Res.* **2000**, *28*, 126-128.
- Pruitt, K. D.; Maglott, D. R. *Nucleic Acids Res.* **2001**, *29*, 137-140.
- Wadelius, M.; Andersson, A. O.; Johansson, J. E.; Wadelius, C.; Rane, E. *Pharmacogenetics* **1999**, *9*, 333-340.
- Adam, R. M.; Borer, J. G.; Williams, J.; Eastham, J. A.; Loughlin, K. R.; Freeman, M. R. *Endocrinology* **1999**, *140*, 5866-5875.
- Montori, V. M.; Smleja, M.; Guyatt, G. H. *Mayo Clin. Proc.* **2000**, *75*, 1284-1288.
- Denenberg, V. H. *Statistics Experimental Design for Behavioral and Biological Researchers*; Wiley-Liss: New York, 1976.
- Rebhan, M.; Chalfin-Caspi, V.; Prilusky, J.; Lancet, D. *Trends Genet.* **1997**, *13*, 163.
- Balroch, A.; Apweiler, R. *Nucleic Acids Res.* **2000**, *28*, 45-48. PR0340227

## mRNA Differential display: application in the discovery of novel pharmacological targets

Xinkang Wang, Robert R. Ruffolo, Jr and Giora Z. Feuerstein

The number of genes in the human genome is estimated at 50 000–100 000. However, only a fraction of these genes are expressed in any one cell. Moreover, the level of gene expression in cells may vary with time, physiological conditions and disease states. This differential gene expression is generally reflected by the different number of mRNA species expressed in a given cell (~15 000 individual mRNA species per cell) at any time point, and changes in relative mRNA levels may have important implications in the development of pathological processes. Therefore, discovery of differentially expressed genes is essential for the understanding of the molecular mechanisms involved in normal and pathological states, as well as providing new insights for discovery of new molecular targets for pharmacological manipulation and drug development. Hence, a number of techniques have been developed to identify genes (with known or unknown sequences and functions) that are differentially expressed in disease states. For example, northern hybridization, RNase protection assay, quantitative reverse transcription and polymerase chain reaction (RT-PCR) have been successfully utilized to identify discordantly expressed known genes. Other techniques, such as differential hybridization and subtractive library screening, have been used successfully for the discovery of differentially expressed genes with known and/or unknown sequences. In the differential hybridization method, a cDNA library is first prepared and then screened using probes that are made from two different sources, for example normal and diseased tissues. Subtractive library screening is carried out on

the basis of the construction of a subtracted cDNA library from different RNA sources, for example normal and diseased cells, of which the identical mRNA species have been removed using hybridization methods. Although these two techniques have proved to be useful in the discovery of differentially expressed genes, they are technically difficult and labour intensive, and require large amounts of mRNA (see Box 1).

Recently, a number of PCR-based methods to uncover differentially expressed genes have been developed; these techniques include (1) mRNA differential display<sup>1</sup>, (2) RNA fingerprinting<sup>2</sup> and (3) arbitrarily primed PCR (Ref. 3). These PCR-based techniques provide some advantages over the conventional methods and have been used successfully for novel gene discovery. In particular, the mRNA differential display methodology has been adopted by a large number of laboratories as an important additional tool that has applications for both *in vitro* and *in vivo* test systems<sup>1,4-7</sup>. An overall strategic approach using this method for drug discovery is outlined in Fig. 1.

### Messenger RNA expression

Messenger RNA is the product of gene expression that encodes for a specific protein. The levels of mRNA in the cell are generally reflected by transcriptional regulation. Following transcription, mRNA is 'matured' by capping the 5'-end, adding the polyadenylation [poly(A)] at the 3'-end, and splicing the intron sequences in eukaryotic cells. Taking advantage of the polyadenylated tail present in most eukaryotic mRNA species, the mRNAs can be reverse-transcribed in the presence of

anchored primers complementary to the 3'-end of mRNAs, such as the use of oligo(dT)<sub>n</sub> (where  $n=12-18$ ) primers. In the technique of mRNA differential display, a set of 3'-anchored primers, such as T<sub>12</sub>MN where M=G, A or C and N=G, A, T or C, are used to prime the reverse transcription reactions.

### Methodology: mRNA differential display

The method of mRNA differential display consists of two basic steps: (1) reverse transcription (RT) using a set of 3'-anchored primers, and (2) PCR amplification of cDNA fragments using arbitrary (upstream) primers and anchored (downstream) primers (Fig. 2).

For the RT reaction, total cellular RNA (DNase treated to eliminate the possibility of genomic DNA contamination) is reverse-transcribed to yield the first strand cDNA primed with T<sub>12</sub>MN oligonucleotides. This RT reaction enables all the mRNA species having a poly(A) tail to be reverse-transcribed. Typically, this RT reaction is divided into four subgroups, each using a different T<sub>12</sub>MN primer with G, A, T or C at the last base of the 3'-end. Because a large number of mRNA species are present in a cell, the division of subgroups for the RT allows a portion of the mRNA species to be displayed, which will increase the resolution of cDNA species after amplification<sup>1</sup>.

Amplification of all the cDNAs is carried out using an upstream arbitrary primer and a downstream anchored primer (identical to the one used for the RT) in the presence of a radioactive nucleotide (Fig. 2). The upstream primer has been optimized to ten bases in length, containing approximately 50% of GC contents<sup>1</sup>. In addition, a relatively low annealing temperature (42°C) is also recommended for the PCR to allow some base mismatches so that a larger number of the amplified mRNA species can be obtained. Using these conditions of amplification, it has been estimated that at

X. Wang,

Investigator,

R. R. Ruffolo, Jr  
Vice President and

Director,

and

G. Z. Feuerstein,

Cardiovascular

Pharmacology

Director,

Division of

Pharmacological

Sciences,

SmithKline Beecham

Pharmaceuticals,

King of Prussia,

PA 19406, USA.

RECEIVED  
NOV 24 2004  
TECH CENTER 1600/2009

### Box 1. Comparison of mRNA differential display with subtractive library screening for novel gene discovery

#### mRNA Differential display

- Key technique is based on RT-PCR
- Very sensitive to detect altered gene expression
- Allows multiple comparison, and monitors both upregulated and downregulated genes
- Relatively reliable to detect the differentially expressed genes; confirmation by other techniques is required
- Rapid to identify a lead probe

#### Subtractive library screening

- Crucial step is subtractive library construction
- Relatively insensitive, especially for those low abundance mRNAs
- Usually compares only unidirectional change
- Very reliable to detect the altered gene expression
- Relatively slow and complicated

least 30–40 upstream primers in combination with the downstream primers will be necessary to amplify every mRNA species present in a given cell<sup>8</sup>.

The amplified cDNA fragments are resolved by electrophoresis and subjected to autoradiographic analysis. By taking advantage of mRNA differential display, multiple samples can be amplified and compared in parallel. As such, differences in gene expression, either upregulated or downregulated, can be identified in specific experimental or pathological conditions or along temporal expression patterns. As shown in Figure 3, the differential display analysis was carried out using cellular RNAs isolated from lipopolysaccharide (LPS)-stimulated and -unstimulated rat aortic vessels<sup>9</sup>.

#### Band recovery

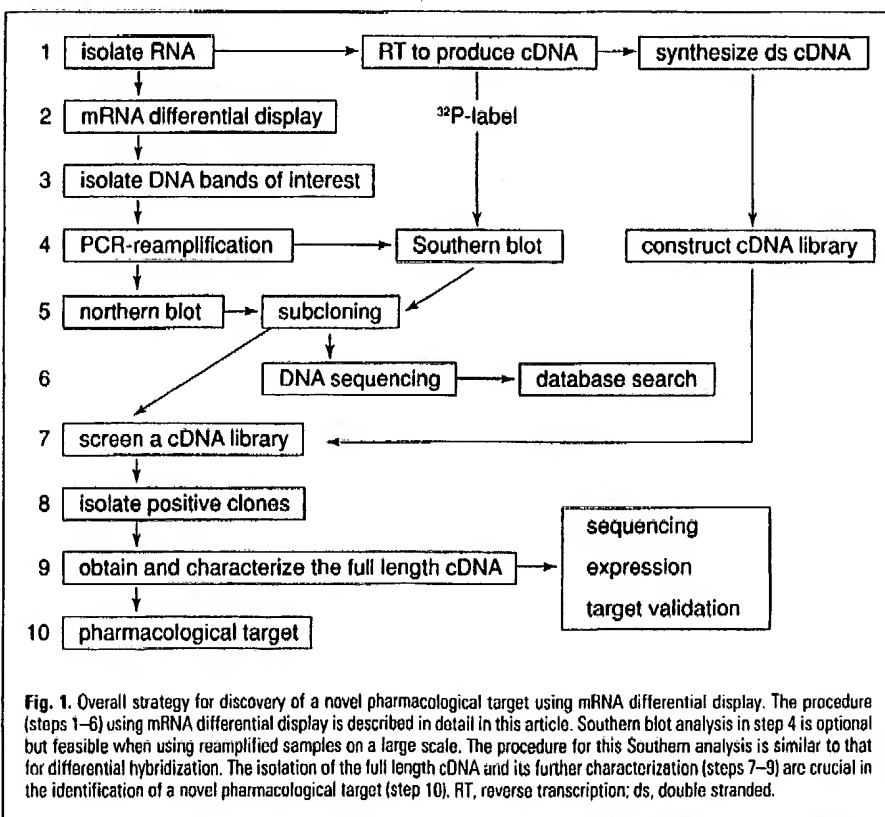
Following mRNA differential display, the bands of interest may be recovered by applying the following three steps: the DNA band is (1) excised from the dried sequencing gel, (2) isolated by extraction procedures, and (3) reamplified using the same sets of primers as in the original PCR (Ref. 1). The recovered DNA band can serve as a probe to confirm mRNA expression by means of northern blot analysis, and/or be subcloned into a vector for further analysis.

#### Confirmation of the differentially expressed genes

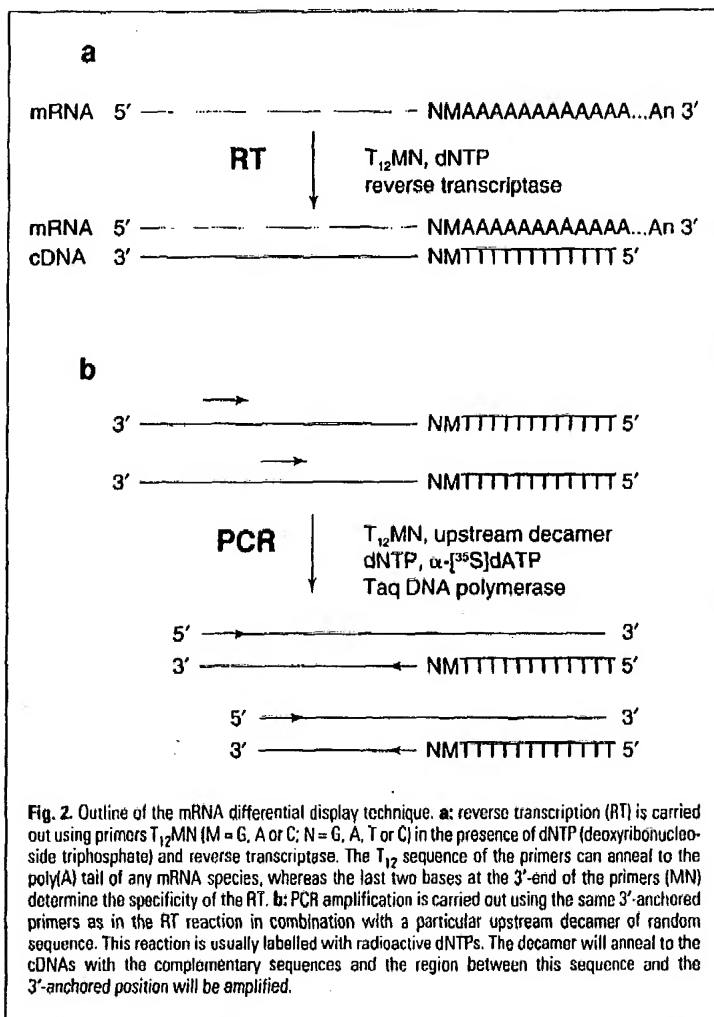
Confirmation of gene expression is one of the crucial steps following mRNA differential display, in as much as a large number of false-positive bands may be present on differential display. A variety of methods to reduce false positives have been utilized in different laboratories; the most commonly used method is

northern blot analysis. Dot blot, quantitative RT-PCR, RNase protection assays and other methods have also been used.

Using two methods, differential display and northern blot analyses, the significant upregulation of mRNA (LPS-7) in response to LPS stimulation in cultured aortic vessels has been confirmed (Figs 3 and 4; see Ref. 9).



**Fig. 1.** Overall strategy for discovery of a novel pharmacological target using mRNA differential display. The procedure (steps 1–6) using mRNA differential display is described in detail in this article. Southern blot analysis in step 4 is optional but feasible when using reamplified samples on a large scale. The procedure for this Southern analysis is similar to that for differential hybridization. The isolation of the full length cDNA and its further characterization (steps 7–9) are crucial in the identification of a novel pharmacological target (step 10). RT, reverse transcription; ds, double stranded.



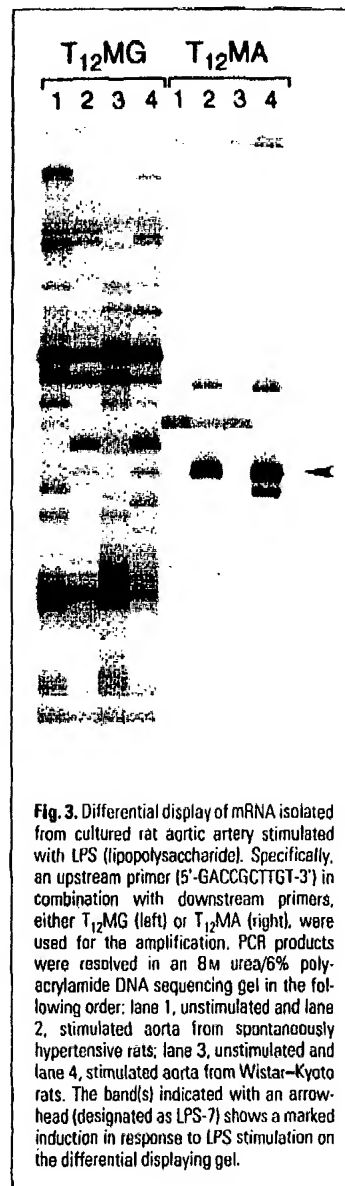
## Identification of the differentially expressed genes

It is fundamental to identify the genes discovered by mRNA differential display. This step relies on the DNA sequencing analysis of the recovered DNA band. Because the primers used for differential display are short and cannot be used successfully for direct sequencing by standard protocols, the differential displayed DNA fragments are typically subcloned into a vector prior to sequencing analysis<sup>1,5</sup>. Recently, direct sequencing of differential display PCR products became feasible (1) on the basis of the use of elongated primers for direct differential display<sup>10,11</sup> or (2) during the reamplification following original differential display method<sup>9</sup>.

Using this sequence information, the identity of the differentially expressed genes can be determined by searching a database, such as GenBank. If the sequence represents an unknown sequence, a cDNA library can be screened using this DNA as a probe in order to obtain the full length cDNA clone.

## Advantages of mRNA differential display

Compared with the conventional methods for the discovery of genes with altered expression in disease states, such as differential hybridization and subtractive library screening, the mRNA differential display technique has several advantages (see Box 1): (1) simplicity in all key techniques (primarily RT-PCR); (2) sensitivity due to PCR amplification;



(3) versatility in detecting genes that are either upregulated or downregulated under various conditions, and the ability to perform a side-by-side comparison of different samples; (4) rapidity in identifying a probe (a cDNA) and confirming the results (e.g. northern blot); (5) small amounts of RNA required; and (6) reproducibility (the displayed bands in general show at least 60-70% identity for different repeats). These characteristics render this technique increasingly popular for the discovery of novel genes.

### Limitations of mRNA differential display

While the differential display technique has significant advantages, some disadvantages in using mRNA differential display must be acknowledged<sup>12</sup>. The major concerns are the high incidence of false positives, and the labour-intensive nature of this procedure for large-scale screening. In addition, the cDNA fragments isolated by this method are typically small, and frequently located in the 3'-untranslated region. Therefore, in order to identify the differentially expressed gene, one may need to screen a cDNA library to isolate the full length cDNA clone. Moreover, in order to observe every differentially expressed gene in the mRNA population, at least 20–25 (and possibly up to 80, see Ref. 13) upstream primers in combination with downstream anchored primers should be used (based upon theoretical calculations)<sup>8</sup>. It is obvious that this technique needs to be refined further in order to be efficiently and widely applied for large-scale searching of altered gene expression in different diseases or under different experimental conditions.

Recently, significant improvements and modifications have been made to the method as originally described<sup>1</sup> in order to overcome some of the existing problems in this technique<sup>14</sup>, e.g. (1) emphasis has been placed on the importance of DNA-free RNA samples and multiple displays of samples; this will reduce the frequency of false positives<sup>15</sup>; (2) longer primers are used, e.g. 18–20 mers, as in RNA-fingerprinting<sup>2</sup>; this not only increases the reproducibility of differential display, but also allows direct sequencing after PCR amplification<sup>10,11</sup>; (3) the application of slot blot has been used to evaluate the bands identified after differential display<sup>16</sup>, or the use of northern blot for affinity capturing of cDNAs (Ref. 17); these methods reduce the labour-intensive nature of this work for large scale screening. Furthermore, (4) the potential hazardous nature of <sup>35</sup>S as a

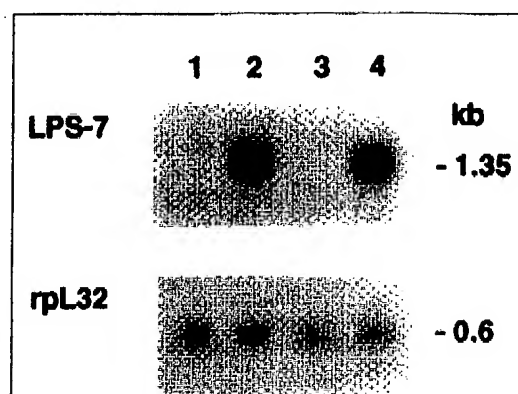
radiolabel for differential display has been noted, and either <sup>32</sup>P or <sup>33</sup>P have been recommended as alternative labels<sup>18,19</sup>.

### Concluding remarks

Differential display of mRNA is one of the most flexible and comprehensive methods available for the detection of differentially expressed genes in the cell. Since its initial description, this technique has been established in many laboratories and applied successfully in the identification of genes using *in vitro* and *in vivo* systems. In addition, other strategies aimed at discovering novel genes are emerging, such as methodology of serial analysis of gene expression (SAGE)<sup>20</sup> and representational difference analysis (RDA)<sup>21</sup>. The application of mRNA differential display, and other techniques, for the isolation of novel genes associated with disease processes will no doubt facilitate the discovery of novel therapeutic targets and/or will help to understand the molecular mechanisms of disease. However, this is the first of many steps (Fig. 1) required in the discovery of a novel pharmacological target, especially given that the function of this factor is most likely unknown. Therefore, further action should be taken to characterize the functions of a particular gene of interest, including isolation of full length cDNA, expression of the gene product for functional study and target validation for the importance of this gene in disease processes.

### References

1. Liang, P. and Pardee, A. B. (1992) *Science* 257, 967–971
2. Welsh, J. et al. (1992) *Nucleic Acids Res.* 20, 4965–4970
3. Sokolov, B. P. and Prockop, D. J. (1994) *Nucleic Acids Res.* 22, 4009–4015
4. Aiello, L. P., Robinson, C. S., Lin, Y. W., Nishio, Y. and King, G. L. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 6231–6235
5. Zimmermann, J. W. and Schultz, R. M. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 5456–5460
6. Utans, U., Liang, P., Wyner, L. R., Karnovsky, M. J. and Russell, M. E. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 6463–6467



**Fig. 4.** Northern analysis of LPS-7 mRNA expression in cultured aorta stimulated with LPS. Total cellular RNA (10 µg/lane, loaded in the order: lane 1, unstimulated and lane 2, stimulated aorta from spontaneously hypertensive rats; lane 3, unstimulated and lane 4, stimulated aorta from Wistar-Kyoto rats) was resolved by electrophoresis, transferred to a nylon membrane, and hybridized to LPS-7 and rpl32 (loading control) cDNA probe, sequentially. The rpl32 mRNA expression was relatively constant in the experimental conditions and therefore used for standardizing the samples loaded in each lane.

7. Wang, X. K. et al. (1995) *Proc. Natl. Acad. Sci. U. S. A.* 92, 11480–11484
8. Bauer, D. et al. (1993) *Nucleic Acids Res.* 21, 4272–4280
9. Wang, X. K. and Feuerstein, C. Z. (1995) *BioTechniques* 18, 448–453
10. Zhao, S., Ooi, S. L. and Pardee, A. B. (1995) *BioTechniques* 18, 842–850
11. Diachenko, L. B., Ledesma, J., Chenchik, A. A. and Siebert, P. D. (1996) *Biochem. Biophys. Res. Commun.* 219, 824–828
12. Debouck, C. (1995) *Curr. Opin. Biotechnol.* 6, 597–599
13. Liang, P. et al. (1995) *Methods Enzymol.* 254, 304–321
14. Liang, P. and Pardee, A. B. (1995) *Curr. Opin. Immunol.* 7, 274–280
15. Liang, P., Averboukh, L. and Pardee, A. B. (1993) *Nucleic Acids Res.* 21, 3269–3275
16. Mou, L., Miller, H., Li, J., Wang, E. and Chalifour, L. (1994) *Biochem. Biophys. Res. Commun.* 199, 564–569
17. Li, F., Barnathan E. S. and Kariko, K. (1994) *Nucleic Acids Res.* 22, 1764–1765
18. Trentmann, S. M., Van der Knaap, E. and Kende, H. (1995) *Science* 267, 1186–1197
19. Tokuyama, Y. and Takeda, J. (1995) *BioTechniques* 18, 424–425
20. Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995) *Science* 270, 484–487
21. Lisitsyn, N., Lisitsyn, N. and Wigler, M. (1993) *Science* 259, 946–951

### Students

Subscribe to *TiPS* at  
a 50% discount

# Gene Families: The Taxonomy of Protein Paralogs and Chimeras

Steven Henikoff,\* Elizabeth A. Greene, Shmuel Pietrokovski, Peer Bork,  
Teresa K. Attwood, Leroy Hood

Ancient duplications and rearrangements of protein-coding segments have resulted in complex gene family relationships. Duplications can be tandem or dispersed and can involve entire coding regions or modules that correspond to folded protein domains. As a result, gene products may acquire new specificities, altered recognition properties, or modified functions. Extreme proliferation of some families within an organism, perhaps at the expense of other families, may correspond to functional innovations during evolution. The underlying processes are still at work, and the large fraction of human and other genomes consisting of transposable elements may be a manifestation of the evolutionary benefits of genomic flexibility.

Linnaeus introduced a universal classification system of living things that was able to organize the enormous complexity of biological relationships. A universal gene classification system presents a similar challenge but with added complexity. If a single gene is likened to an individual, then the collection of genes sharing common ancestry, typically performing the same role in different organisms, would be analogous to a species. Genes that are related in this way are commonly referred to as "orthologs" (1). Higher levels of gene or protein classification, such as families, subfamilies, and superfamilies, create a hierarchy in molecular taxonomy (2). Just what constitutes gene classification criteria can be uncertain in practice. This situation is made much more uncertain by the existence of nonorthologous relationships. Multiple proteins resulting from gene duplications within an organism are termed "paralogs." Paralogous relationships have been known for several decades:  $\alpha$ -globin,  $\beta$ -globin, and myoglobin are classical examples of paralogs that arose from duplications of ancestral globin genes in the vertebrate lineage (3). In recent years, with the explosive increase in available sequence data, we have become aware of the richness of paralogous relationships in all organisms. We now realize that protein building blocks, or "modules," have duplicated and evolved in complex ways

through a variety of gene-rearrangement mechanisms (4). As a result, composite proteins consisting of multiple modules ("chimeras") constitute a large proportion of the protein complement of an organism. The complexity that results from so many paralogous and chimeric relationships presents a daunting challenge for classification. Meeting the challenge unites sequence with biological information.

Like taxa, which reflect common ancestry but can also be used to infer common function, gene families have been of tremendous importance for understanding gene and protein function. Nearly all biological disciplines have profited from discoveries of family relationships. Such discoveries have emphasized the importance of model systems in biology. For example, the sequencing of *Drosophila Ultrabithorax* and *Antennapedia* selector genes controlling segment identity delineated a shared homeobox module; this led to the discovery and intense study of related HOX genes in vertebrates and other organisms that are thought to play key roles in determining developmental fates (5). This example illustrates an increasingly popular paradigm in molecular genetics: Rather than proceeding from a phenotype to the isolation of a new gene, an investigator begins with the sequence of a key gene and searches for homologous genes in an organism of interest, preferably by scrutinizing the sequence databanks (6). Experimental data accumulated for the homologous (orthologous or paralogous) gene, when integrated with insights from gene family relationships, can accelerate our understanding of biological processes and our ability to rationally engineer genes.

Not just functional, but also structural inferences made from protein sequence alignments have been valuable to biologists. When a structure is known for one sequence, and another can be aligned with

it, the unknown backbone structure can be predicted with confidence. In the case of homeoboxes, the high level of inferred structural similarity has guided site-directed modification of this DNA-binding domain for homeoboxes other than the structural archetype, and this situation holds for ~30% of known protein sequences (7).

## Motifs, Modules, and Chimeras

The smallest sequence units of protein families are termed "motifs," which are identified as highly similar regions in alignments of protein segments (8). Motifs can be as simple as the hexamer repeat unit that forms a left-handed parallel  $\beta$ -helix found in uridine 5'-diphosphate (UDP)-*N*-acetylglucosamine acyltransferase (9). Motifs are widely used to identify functional regions of proteins and, where they share common ancestry, are useful for family classification. The C<sub>2</sub>H<sub>2</sub> zinc finger DNA-binding motif, which is illustrated in the accompanying chart, defines the largest known family. By virtue of forming a contiguous independently folded structure, the finger is itself a module, whose small size of 21 to 26 amino acids is attributable to a zinc cation, which holds together two cysteine and two histidine residues from either end of the module. The larger homeobox module consists of a ~60-amino acid motif also involved in binding DNA. More typically, modules consist of multiple motifs, which form the structural core of proteins. Motifs contributing to a structural core can be widely separated within the primary sequence, as illustrated by the "HIGH" and "KMSKS" motifs of the Class I aminoacyl tRNA synthetases, which are hundreds of amino acids apart (10). Enzyme active site residues, which are usually highly conserved, are often found within motifs.

Motifs may reflect either common ancestry or convergence from independent origins. In either case, identification of motifs can be important for drawing structural and functional inferences. For example, the common "P-loop" motif is present in nucleotide-binding domains from families as diverse as kinesin motor proteins and adenosine 5'-triphosphate (ATP)-binding cassette (ABC) transporters, which are depicted in the accompanying chart. Despite the

S. Henikoff is at the Fred Hutchinson Cancer Research Center and Howard Hughes Medical Institute, Seattle, WA 98109-1024, USA. E. A. Greene and S. Pietrokovski are at the Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA. P. Bork is at the European Molecular Biology Laboratory, 69012 Heidelberg, Germany, and Max-Deibueck-Center for Molecular Medicine, 13122 Berlin-Buch, Germany. T. K. Attwood is in the Department of Biochemistry and Molecular Biology, University College London, London WC1E 6BT, UK. L. Hood is in the Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195, USA.

\*To whom correspondence should be addressed.



lack of a known structure for any ATP-binding cassette, the presence of a P-loop predicts the site of ATP binding in the transporter complex.

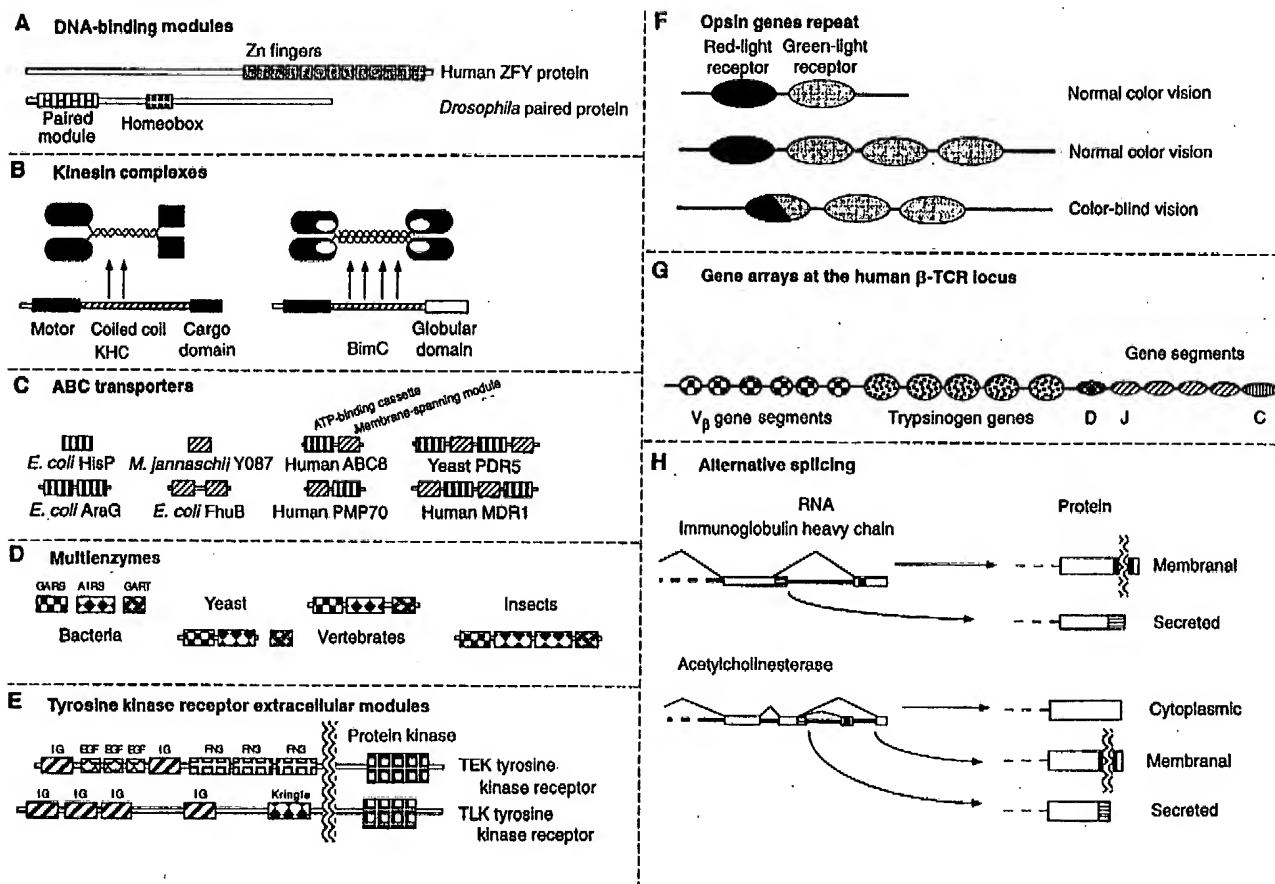
Modules are composed of single or multiple motifs. As the fundamental units of protein structure and function, modules are most useful for protein classification. Modules frequently display different connectivity relationships (Fig. 1, A to F), as illustrated by the kinesins and ABC transporters. The kinesin motor domain can be at either end of a polypeptide chain that includes a coiled-coil region and a cargo domain (11). ABC transporters are four-domain proteins consisting of two unrelated modules, a pair of ATP-binding cassettes, and a pair of integral membrane modules, which can be connected in different ways (12) (Fig. 1C).

## Dispersal of Protein Building Blocks

Family relationships evolve over long periods of time by speciation and by sequence duplications fixed in genomes. Even the most recently evolved family relationships are still so ancient that the events that gave rise to paralogs and chimeras in modern genomes cannot be directly observed. However, enough is known about genomic-rearrangement mechanisms that some inferences can be drawn. Chromosomes evolve by transposition of mobile elements; by gross rearrangements such as inversions, translocations, deletions, and duplications; by homologous recombination; and by slippage of DNA polymerases during replication. It is likely that all of these mechanisms have contributed to the proliferation and dispers-

al of protein building blocks. Modules present in larger proteins, including homeobox modules, might have dispersed by transposition. Tandemly repeated modules, including the  $C_2H_2$  zinc fingers and many examples of extracellular modules, most likely arose by recombinational mechanisms, such as unequal crossing-over and gene conversion (Fig. 1, A and E).

Multiple eukaryotic biosynthetic enzymes, especially those in the purine and pyrimidine pathways, are sometimes found together within a single polypeptide, unlike their separately encoded bacterial orthologs (13). For example, vertebrates have a multienzyme polypeptide for GAR synthetase, AIR synthetase, and GAR transformylase (GARS-AIRS-GART) (14). In insects, the polypeptide appears as GARS-(AIRS)<sub>2</sub>-GART; in yeast, GARS-AIRS is encoded



**Fig. 1.** Schematic representations of various building block arrangements described in the text. (A) Simple building blocks in DNA-binding proteins. The human ZFY protein contains 13 tandemly repeated zinc finger modules, and the *Drosophila* paired protein contains a paired box and a homeobox. (B) Subfamily relationships as predictors of quaternary structure: dimeric kinesin heavy chain (KHC) and tetrameric BimC protein complexes. (C) ABC transporters display different connectivities of two subunit pairs. Other examples of circular permutation have been recently reviewed

(54). (D) Organism-specific fusion and duplication of purine biosynthetic pathway orthologs to GARS, AIRS, and GART. (E) Diverse modules are found in the extracellular portion of protein tyrosine kinases. (F) Humans are polymorphic for duplications and deletions within the opsin tandem cluster of long-wavelength genes. (G) T cell receptor (TCR) genes are interrupted by clusters of  $\beta$ -trypsinogen genes. (H) Alternative processing produces membrane-bound, secreted or intracellular forms of antibodies (or both), and acetylcholinesterases.



separately from GART; and in bacteria, GARS, AIRS, and GART are all encoded separately (Fig. 1D). The sites of fusion may correspond to introns, suggesting that chromosomal rearrangements have fused transcription units within introns. In other cases, fusions might have occurred in exons, or intron loss might have erased evidence of intron-mediated fusion (15). Regardless of mechanism, the fusion of transcription units is likely to have contributed to combining of protein building blocks in both eukaryotes and prokaryotes.

The mechanisms that gave rise to the dispersal of paralogous proteins within genomes are also diverse and frequently uncertain. The rhodopsin-like guanosine 5'-triphosphate (GTP)-binding protein (G protein)-coupled receptors illustrate multiple dispersal patterns (16). This family includes hormone, neurotransmitter, light, and olfactory receptors that are distinguished from one another by both sequence and functional differences. Remarkably, there are several hundred human olfactory receptor (OR) genes present in a dozen or so tandem clusters on several chromosomes (17). A cluster of three OR genes and an OR pseudogene fused to a different OR gene is thought to have arisen from disparate events, including recombinations between repeats flanking OR genes and a fusion by nonhomologous deletion (18).

Tandem gene clusters are sometimes interrupted by paralogous members of other gene families. For example, intercalated between repeated coding elements of the human  $\beta$  T cell receptor (TCR) locus are five trypsinogen genes in inverted orientation (19) (Fig. 1F). This complex arrangement of genes is likely to be of functional significance, as it is also found in mice and chickens.

Many paralogous relationships might be the consequence of whole-genome duplications. Ancient tetraploidization events in eukaryotes have been obscured by subsequent divergence, interchromosomal duplications, and other rearrangements but can be detected by careful analysis of genomic sequence. For example, it has been proposed that the *Saccharomyces cerevisiae* genome underwent a whole-genome duplication, and that 13% of *Saccharomyces cerevisiae* genes trace their lineage to this event (20). Tetraploidization events are common among higher plants; for example, the wheat genome consists of three copies of an ancestral grass genome. The human genome is thought to be the product of multiple tetraploidization events that occurred during chordate evolution (5). As a result, we have four copies of many genes or gene families, including

four *HOX* gene clusters comparable to a single set of *HOX* genes in invertebrates. Enough time has passed since these putative tetraploidization events that vertebrate *HOX* genes have acquired distinguishable functions.

### Selection for Diversity

The acquisition of a new specificity or a modified function after a gene-duplication event is often detectable by protein sequence comparison. For example,  $\alpha$ -globins are more closely related to one another than they are to any  $\beta$ -globin. Maintenance of an acquired function over long evolutionary intervals can contribute greatly to the understanding of gene specificity. For example, sequence differences are sufficient to distinguish among tRNA synthetases that charge different amino acids, even though they belong to the same ancestral family (21). The kinesin motor domains provide another example, where relationships within a family are predictors for quaternary structural features: BimC motor domains are found in bipolar complexes, rather than in asymmetric complexes characteristic of other kinesin motors (22) (Fig. 1B). Comparisons should be interpreted with caution, especially when sequences from very distant organisms are compared; apparent subfamily relationships will not always reflect shared function. Furthermore, similar functions can arise in separate subfamilies. For example, among the ABC transporters, iron uptake is a function of members of two distinct subfamilies (23).

Relatively recent duplication events are sometimes responsible for diversity in molecular recognition. Tandem duplication of immunoglobulin (Ig) and TCR variable, joining, and diversity gene segments is the prototypical example, and special mechanisms of somatic DNA rearrangement and mutation further diversify antibody and TCR specificity. Among the rhodopsin-like G protein-coupled receptors, different olfactory receptors are thought to recognize different odorants, and different opsins are stimulated by different wavelengths of light. Long- and short-wavelength opsin genes diverged from one another early in vertebrate evolution (24). The opsins of the human visual system are present in a cluster on the X chromosome, with the long-wavelength opsins, sensitive to red and green light, constituting a tandem repeat with 98% sequence identity (Fig. 1F). Remarkably, the number of long-wavelength genes is polymorphic, a consequence of unequal crossing-over events that have occurred during human evolution. People with "normal" vision have a single red gene and one to three green genes. People who are red-green colorblind have lost a

long-wavelength gene through a fusion of red and green tandem copies.

The products of gene duplication can act combinatorially and so further increase diversity. A response to a single antigen generally stimulates the proliferation of different B cells, each expressing a single antibody; the combination of different light and heavy chains provides heightened specificity to antigen. For olfaction, the stimulation of multiple olfactory receptors by their different odorants allows complex mixtures to be recognized. Our ability to recognize a full spectrum of colors with only three types of opsins is another example of the integration of multiple sensory inputs that have originated from duplicated building blocks.

Duplication of building blocks within a protein also results in generation of diversity during evolution. Each  $C_2H_2$  zinc finger in a DNA-binding protein can recognize a 3-base pair motif, and in combination, multiple zinc fingers can mediate the binding to more complex DNA recognition sites (25). Combinatorial recognition by tandem zinc fingers has been exploited by researchers for designing new DNA-binding proteins (26). Combinations of unrelated modules have also broadened the spectrum of DNA-binding recognition, such as the presence of a paired box and a homeobox module in proteins related to *Drosophila paired* (27) (Fig. 1A). Extracellular proteins are notable for containing combinations of multicopy tandem arrays of different modules. The extracellular portion of the receptor tyrosine-specific class of protein kinases contains an astonishing variety of modules representing different families. For example, *trk*-like kinases have one kringle and four Ig modules, whereas *tek*-related proteins have three fibronectin III, three epidermal growth factor (EGF), and two Ig modules in their extracellular  $NH_2$ -terminal portions (28) (Fig. 1E). These extracellular modules can acquire diverse functions in different proteins. For example, some EGF modules bind to specific receptors, whereas others mediate interactions through calcium binding; the latter sometimes form long, rodlike structures composed of tandem module arrays (29).

Unlike germ-line processes that recombine gene segments during evolution, alternative messenger RNA (mRNA) processing can increase the diversity of proteins in the soma. For example, an alternative polyadenylation site within an intron of the Ig heavy-chain gene allows a switch from the synthesis of a membrane-bound receptor to a secreted antibody (30) (Fig. 1H). Acetylcholinesterase provides an example of alternative 3' splice site selection accomplishing a comparable task; the choice of one terminal



exon leads to the synthesis of a glycoprotein membrane anchor, the choice of the other to a cytoplasmic form, and lack of splicing to a secreted form of the enzyme (31).

### Why Are Some Families So Large?

The accompanying chart provides information on the distribution of selected building blocks in model organisms. For organisms with completely determined genomic sequences, we can ask why some families are more successful than others. In *Escherichia coli*, the ABC transporters are the most common proteins encoded; this might reflect a flexible diet, which requires the uptake of diverse nutrients (12). It is likely that the much smaller number of ABC transporters in *Mycoplasma genitalium* and *Methanococcus jannaschii* reflect more limited diets. In general, paralogs account for half of all *E. coli* genes (32), which is high compared to the fractions found for smaller bacterial genomes, such as *Haemophilus influenzae*, where one-third of all genes are paralogs (33, 34). Much of this difference is attributable to the more diverse nutritional and metabolic requirements of *E. coli* (34).

For organisms that have not yet been fully sequenced, it is necessary to extrapolate from samples of available sequences. For example, on the basis of finding only eight homeobox genes in *S. cerevisiae*, extrapolation predicts about 20 each in flies and worms, which are estimated to have two to three times as many genes (see accompanying chart). The fact that there are already about 60 genes reported in each of these two complex multicellular organisms demonstrates that homeobox genes have more successfully proliferated in animals than in a yeast. Although the number from

*Drosophila melanogaster* is based on only ~10% of its genome, we predict that most of its homeobox genes have already been identified, and the final number will not be much greater than the number in *Caenorhabditis elegans* (which has nearly the same sized genome, ~70% of which is already sequenced). Such disproportionate representation of particular families is both a manifestation of their intense interest to researchers and of the ability to obtain these members by hybridization and amplification methods. Not all modules are as amenable to this approach as are the homeoboxes, which are especially highly conserved; to an increasing extent, partial complementary DNA (cDNA) sequencing projects are being used to identify coding sequences for gene families of interest (35). Many other gene families, such as the globins and the immunoglobulins, are disproportionately represented in collections of human sequences because they are important for human health (Table 1).

Even for the whole-genomic sequences that are currently available, the final size of known families is uncertain. Distant homologs may lie just beyond the horizon of current homology-detection methods. However, the introduction of improved methodology continues unabated, and this has led to the discovery of new family members and interfamily relationships. Moreover, the increasing size of a family can be exploited by multiple sequence-based methods to identify additional members (36). For example, 12 years ago, the similarity between opsin genes from human and fly was barely at the level of detection (37), yet today, the opsins are recognized as a closely related cluster within the rhodopsin-like G protein-coupled receptors (see accompanying chart). Most importantly, the accumulation of experimental evidence concerning gene or protein function

or protein structure will provide insights that can be used to deduce possible family relationships that would not be compelling by sequence comparison methods alone.

### Phylogenetic Distribution of Families

Size of a family within an organism is only one measure of success. Another is presence of a family in diverse organisms. Some families are successful at both, such as the ABC transporter family, which is not only one of the largest families overall (Table 1), but also appears to be present in all organisms. Most other families that are so widely distributed show much less proliferation within organisms. These include metabolic enzymes and components of the translational apparatus, which have only a few close paralogs (38). These families show a similar distribution to that of the GARS module in the table of the accompanying chart (39).

The chymotrypsin family of serine proteases is notable in being both ancient and large (Table 1), but the extreme proliferation appears to be confined to eukaryotes; only rarely are family members found in bacteria. This raises the possibility that other families that appear to be confined to certain branches of the tree of life are actually more ancient, but that they have simply become extinct in other lineages, or that a relationship has gone undetected. The latter is the case for eukaryotic tubulin and bacterial FtsZ, both of which use GTP for polymerization to form similar intracellular fibers and are believed to be ancestrally related (40). This relationship was not detected by pairwise sequence comparisons, but rather by recognition of a tubulin motif in FtsZ. Potentially homologous proteins have also been identified by structure determination, such as the detection of similar folds for kinesin and myosin motor proteins (41).

Given the extreme uncertainty in tracing the birth of a family, we nevertheless recognize that some families have proliferated to a remarkable extent in certain phyla. GAL4 transcriptional regulators, one of the largest families in yeast, have been found only in fungi (see accompanying chart). The EGF module, present in about 1% of human proteins, has been described only in animals (Table 1). The Ig module, which is found in more than 200 proteins in addition to all of the immune receptors (antibodies, TCRs, class I and II families of the major histocompatibility complex), is involved in diverse cell surface recognition phenomena in multicellular organisms (42). The Ig module has also successfully proliferated within proteins: A total of 244 copies of Ig and

**Table 1.** The largest protein families. The sources for these numbers of modules are Pfam (PF) or Prints (PR). GPCR, G protein-coupled receptor; LDL, low density lipoprotein.

Family	Source	Modules in SwissProt	Found where?
C <sub>2</sub> H <sub>2</sub> zinc fingers	PF00096	1826	Eukaryotes, archaea
Immunoglobulin module	PF00047	1351	Animals
Protein (Ser/Thr/Tyr) kinases	PF00069	928	All kingdoms
EGF-like domain	PF00008	854	Animals
EF-hand (Ca binding)	PF00036	790	Animals
Globins	PF00042	699	Eukaryotes, bacteria
GPCR-rhodopsin	PF00001	597	Animals
Fibronectin type III	PF00041	514	Eukaryotes, bacteria
Chymotrypsins	PR00722	464	Eukaryotes, bacteria
Homeodomain	PF00046	453	Eukaryotes
ABC cassette	PF00005	373	All kingdoms
Sushi domain	PF00084	343	Animals
RNA-binding domain	PF00076	331	Eukaryotes
Ankrin repeat	PF00023	330	Eukaryotes
RuBisCo large subunit	PF00016	319	Plants, bacteria
LDL receptor A	PF00057	309	Animals



distantly related fibronectin III modules account for most of the 30,000-residue muscle titin protein (43). The success of the ~100-amino acid Ig module is attributable to its potential to undergo diversification in the presence of a highly conserved structural framework, its protease resistance in the folded form, and its ability to readily form homo- and heterodimers through multiple interacting surfaces, so that it is especially suitable for mediating cell-cell interactions.

Proliferation of one family might have occurred at the expense of others. The distribution of protein kinases is suggestive, in that the family consisting of serine-, threonine-, and tyrosine-specific enzymes is hugely successful only in eukaryotes, but is poorly represented in bacteria (see accompanying chart). Conversely, the family of histidine-specific protein kinases is highly successful in *E. coli* and other bacteria, but is relatively rare in eukaryotes. In such situations, we must also consider the possibility that these families are recent arrivals in some organisms, having been transferred horizontally between kingdoms. Horizontal transfers are difficult to document unless there are conspicuous anomalies evident from molecular phylogenetic analyses. Such anomalies have indicated numerous horizontal transfers of mariner transposases between diverse animals (44), as well as transfer of the fibronectin III module from a eukaryote to a bacterium (45).

The establishment, proliferation, or extinction of a protein family in a lineage may coincide with a functional innovation during evolution. For example, actins, tubulins, and motors such as kinesins are found only where there is a cytoskeleton, as though the evolution of these proteins was coordinate with the appearance of the cytoskeleton in eukaryotes. In bacteria,  $\sigma$  factors regulate transcriptional initiation, in contrast to eukaryotes and archaea, which use a different system (46). This difference suggests that either the  $\sigma$  factor system coincided with the appearance of bacteria or that it was lost in the eukaryotic-archaea lineage.

### Interspersed Genomewide Repeats

Analysis of whole-genomic sequences definitively demonstrates that coding regions of genes dominate the prokaryotic genome (38). In contrast, complex eukaryotic genomes are dominated by noncoding sequences. Families of repeats derived from transposable elements constitute a major portion of these eukaryotic genomes, far exceeding exons in the proportion of the genome devoted to them (47, 48). Transposition can occur by reverse transcription of an

**Table 2.** Content of long contiguous stretches of DNA sequence in selected human and mouse gene regions. Data are from the Leroy Hood laboratory.

Region	Contig length (bp)	GC (%)	mRNA (%)	Interspersed repeats (%)	Line1 (%)	Alu or B1/B2 (SINES) (%)
Human TCR $\alpha$	1,071,650	40	4.0	35	16	8
Mouse TCR $\alpha$	228,654	41	1.5	33	22	2.4
Human TCR $\beta$	684,973	42	4.6	30	14	5
Human TCR on chromosome 9	216,293	41	1.7	45	23	9
Mouse TCR $\beta$	700,960	40	3.8	43	32	2
Human MHC class III	299,287	52	16.8	30.5	6.7	17

RNA intermediate or by excision and reintegration of DNA itself (DNA transposition). These elements fall into four categories: short interspersed nuclear elements (SINEs), long dispersed nuclear elements (LINEs), long-terminal repeat (LTR) retrovirus-like elements, and DNA transposons (Fig. 2). In the human, there are ~1,100,000 Alu sequences (a SINE) and 590,000 Line1 sequences (a LINE). It is impressive that Line1 occupies an order of magnitude more of our genome than all of our gene-coding sequences combined. Furthermore, with improved techniques for identifying degraded repeat sequences, perhaps 50% of our genome and an even higher fraction of the mouse genome will be found to consist of genomewide repeats. Much of the nonassigned genome sequences might be composed of interspersed repeats degraded to the point that they are no longer recognizable.

Vertebrate chromosomes have large-scale mosaic structures, or isochores, often with distinct ratios of G+C nucleotides, repeat content, and gene density (49). The human contigs in Table 2 represent high (class II major histocompatibility locus)-, medium (TCR)-, and low (metabolic glutamate receptor 8)-gene density regions. Low-gene density loci are A+T- and Line1-rich, whereas high-gene density loci are G+C- and Alu-rich (47, 49). The A+T-rich isochores, in general, contain longer genes.

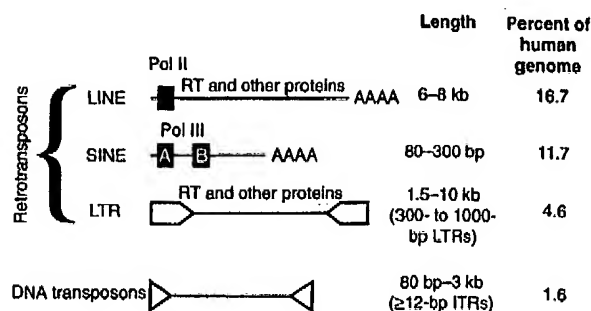
The repeats may have at least three important functional and evolutionary roles. First, some may evolve to become

the regulatory regions of genes expressed in a tissue-specific manner (50). Second, repeats play an important role in refashioning the genomic architecture by facilitating homologous recombination, translocations, and perhaps gene conversions. And third, repeats have been implicated in epigenetic phenomena, such as parental imprinting and position-effect variegation (51). Because the ages of repeats can be determined by species comparisons, they can serve as valuable time markers for unraveling the complexities of molecular archaeology in complex gene loci such as the TCR genes.

### Prospects

There is good news and bad news for gene taxonomists. The good news is that the number of identified protein families has been increasing only slowly with the rapid increase in new sequence data and is expected to level off. The bad news is that family relationships are so complex that we cannot use any simple hierarchical scheme to make the data easily understandable. Nevertheless, as more is learned from model organisms about individual modules, their presence in any protein of interest adds potential insight into its function and guides experiments, which is good news for biologists. Gene taxonomists have learned by now to cope with complexity in family relationships, and currently several classification systems are used to construct the different databases

**Fig. 2.** Schematic representation of the types of transposable elements that have produced high-copy number human interspersed repeats. The shaded boxes denote internal promoter sites; names inside the bracket indicate that only autonomous elements code for these proteins. LTR, long-terminal repeat; ITR, inverted-terminal repeat; RT, reverse transcriptase. [Adapted from (47) on the basis of 7051 kb of human sequence]



listed in the accompanying chart. In fact, the task of classification is made easier for gene taxonomists than for Linnaean taxonomists because sequence similarity is a precisely defined metric for establishing relatedness. This metric makes possible automated and computer-assisted classifications of genes. Much more difficult is the task of enriching the databases of genes and families with insights obtained from experiments.

To some extent, computer-based tools can be applied to the task of connecting genes and families with information about them. Organism-specific databases and retrieval tools such as the National Center for Biotechnology Information's Entrez allow biologists to rapidly obtain needed information from the World Wide Web. However, insight cannot be automated, and computer-based tools that go beyond sophisticated retrieval methods may not be the solution. One problem is that generalized databases are too constraining to allow more than minimal documentation of individual protein families. Another problem is that the literature pertaining to a single family can be so vast that only an expert devoted to that family can master it. Fortunately, a number of biologists interested in particular families have begun to exploit the Web to provide the kind of rich information that can be used to gain insight into function. At a single family Web site, participation can be distributed among multiple laboratories, and information can be continually updated and integrated (52). Furthermore, new Web sites are developed on the basis of existing sites. There are currently five Web sites dedicated to different nuclear hormone receptors spawned from the Nuclear Receptor Resource, and the Myosin Web site was spawned from the Kinesin Web site (53). An organized effort to develop such sites is in progress (see <http://proweb.org> for information on participating).

We have focused here and in the accompanying chart primarily on large and well-studied families. But to truly understand a biological system, we will need to understand the interaction of all individual components. Some of these components will not be immediately classifiable. Eventually, detectable homologs for most of these "orphans" will be discovered in genome-sequencing projects. As a result, new family relationships will become delineated that are useful for identifying critical regions and guiding experimen-

tal work. This situation is most evident in an organism such as *M. jannaschii*, for which a large fraction of proteins are as yet unclassified orphans, but to a lesser extent it is true for all major phyla. The identification and classification of new protein families and the deep insights that result should continue well into the next millennium.

## REFERENCES AND NOTES

1. W. Fitch, *Syst. Zool.* **19**, 99 (1970). Orthologs can only be determined definitively with a complete inventory of the genes in an organism. See R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).
2. We use the term "family" generically to describe any collection of genes or proteins that are presumed to share common ancestry.
3. V. M. Ingram, *Hemoglobins In Genetics and Evolution* (Columbia Univ. Press, New York, 1963).
4. Modules are contiguous in sequence, whereas structural domains are independently folded units that need not be contiguous [L. Patthy, *Cell* **41**, 657 (1985); S. Henikoff, J. C. Wallace, J. P. Brown, *Methods Enzymol.* **183**, 111 (1990); P. Green et al., *Science* **259**, 1711 (1993); R. F. Doolittle, *Annu. Rev. Biochem.* **64**, 287 (1995)].
5. W. J. Gehring and Y. Hironaka, *Annu. Rev. Genet.* **20**, 147 (1986). F. H. Ruddle et al., *ibid.* **28**, 423 (1994).
6. R. F. Doolittle, *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences* (University Science Books, Mill Valley, CA, 1987).
7. C. Orengo, *Curr. Opin. Struct. Biol.* **4**, 429 (1994); R. Schneider, A. de Ruvo, C. Sander, *Nucleic Acids Res.* **25**, 226 (1997).
8. The term "motif" has different interpretations. See P. Bork and E. V. Koonin, *Curr. Opin. Struct. Biol.* **6**, 366 (1996).
9. C. R. H. Raetz and S. L. Roderick, *Science* **270**, 997 (1995).
10. P. Schimmel, *Trends Biochem. Sci.* **16**, 1 (1991).
11. J. D. Moore and S. A. Endow, *Bioessays* **18**, 207 (1996).
12. M. Dean and R. Allikmets, *Curr. Opin. Genet. Dev.* **5**, 779 (1995).
13. J. N. Davidson et al., *Bioessays* **15**, 157 (1993); J. N. Davidson and M. L. Peterson, *Trends Genet.* **13**, 281 (1997).
14. GARS (glycinamide ribonucleotide synthetase), AIRS (aminimidazole ribonucleotide synthetase), and GART (glycinamide ribonucleotide transformylase).
15. A. Rzhetsky et al., *Proc. Natl. Acad. Sci. U.S.A.* **94**, 6820 (1997); S. J. de Souza et al., *ibid.* **93**, 14632 (1996).
16. G protein-coupled receptors are a "clan," which includes proteins that may not be ancestrally related to rhodopsin.
17. N. Ben-Arie et al., *Hum. Mol. Genet.* **3**, 229 (1993). R. R. Reed, *Cold Spring Harbor Symp. Quant. Biol.* **57**, 501 (1992).
18. G. Glusman et al., *Genomics* **37**, 147 (1996).
19. L. Rowen, B. F. Koop, L. Hood, *Science* **272**, 1755 (1996).
20. K. H. Wolfe and D. C. Shields, *Nature* **387**, 708 (1997).
21. R. Wetzel, *J. Mol. Evol.* **40**, 545 (1995).
22. A. S. Kashina et al., *Nature* **378**, 270 (1996).
23. A. Angerer, S. Gaisner, V. Braun, *J. Bacteriol.* **172**, 572 (1990).
24. J. Nathans et al., *Annu. Rev. Genet.* **26**, 403 (1992).
25. Y. Choo and A. Klug, *Curr. Opin. Struct. Biol.* **7**, 117 (1997).
26. ———, *Curr. Opin. Biotechnol.* **6**, 431 (1995).
27. D. Bopp et al., *Cell* **47**, 1033 (1986).
28. P. Maslakowski and R. D. Carroll, *J. Biol. Chem.* **267**, 26181 (1992); J. Partanen et al., *Mol. Cell. Biol.* **12**, 1698 (1992).
29. P. Bork et al., *Q. Rev. Biophys.* **29**, 119 (1996).
30. J. Rogers et al., *Cell* **20**, 303 (1980).
31. Y. Li, S. Camp, P. Taylor, *J. Biol. Chem.* **268**, 6790 (1993).
32. B. Labedan and M. Riley, *Mol. Biol. Evol.* **12**, 980 (1995).
33. S. E. Brenner et al., *Nature* **378**, 140 (1995).
34. R. L. Tatusov et al., *Curr. Biol.* **6**, 279 (1996).
35. M. D. Adams et al., *Science* **252**, 1651 (1991).
36. R. F. Doolittle, Ed., *Methods Enzymol.* **266** (1996).
37. J. E. O'Tousa et al., *Cell* **40**, 839 (1985).
38. R. L. Tatusov et al., in (7).
39. For the table in the accompanying chart, organism-specific counts were obtained for C<sub>2</sub>H<sub>2</sub> zinc fingers (Pfam PF00096), homeodomains (Blocks BL00027), LysR transcription regulators (BL00044), TATA-binding protein repeat (BL00351), 7TM rhodopsin-like receptors (Prints GPCRHHODOPSIN), kinesin motors (BL00411), ATP-binding cassette (BL00211), DEAD/H helicases (PF00270), AAA modules (BL00674), hsp60s (BL00296), and hsp20s (BL01031) by MAST searches [T. L. Bailey and M. Gribskov, *J. Comp. Biol.* **4**, 45 (1997)] of OWL version 29.3 by use of position-specific scoring matrices from local multiple alignments [J. G. Henikoff and S. Henikoff, *Comput. Appl. Biosci.* **12**, 135 (1996)]. For GAL4 transcription regulators, Ser-, Thr-, Tyr-specific kinases, His-specific kinases, kringle extracellular domain, WW intracellular domain, BRCA1 COOH-terminal domain, and Calponin homology domain, profiles were constructed from multiple alignments and used to search an exhaustive protein database at the European Molecular Biology Laboratory, Heidelberg, Germany, with exclusion of redundant entries [P. Bork and T. J. Gibson, *Methods Enzymol.* **266**, 162 (1996)].
40. H. P. Erickson, *Cell* **80**, 367 (1995).
41. F. J. Kull et al., *Nature* **380**, 550 (1996).
42. T. Hunkapiller and L. Hood, *Adv. Immunol.* **44**, 1 (1989).
43. S. Labelle and B. Kolmerer, *Science* **270**, 293 (1995).
44. H. M. Robertson, *Nature* **362**, 241 (1993); *J. Hered.* **88**, 195 (1997); ——— et al., *Nature Genet.* **12**, 360 (1996).
45. P. Bork and R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8990 (1992).
46. D. Langer et al., *ibid.* **92**, 5768 (1995).
47. A. F. A. Smit, *Curr. Opin. Genet. Dev.* **6**, 743 (1996).
48. P. SanMiguel et al., *Science* **274**, 765 (1996); J. A. Yoder, C. P. Walsh, T. H. Bastor, *Trends Genet.* **13**, 335 (1997).
49. G. Bernardi, *Annu. Rev. Genet.* **29**, 445 (1995).
50. J. Brosius, *Science* **251**, 753 (1991); S. E. White, L. F. Habera, S. R. Wessler, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 11792 (1994); R. J. Britten, *ibid.* **93**, 9374 (1996).
51. S. Henikoff and M. A. Matzke, *Trends Genet.* **13**, 293 (1997); D. P. Barlow, *Science* **260**, 309 (1993).
52. S. Henikoff, S. A. Endow, E. A. Greene, *Trends Biochem. Sci.* **21**, 444 (1996).
53. Nuclear Receptor Resource, <http://nrr.georgetown.edu/NRR/NRR.html>; Kinesin Home Page, <http://proweb.org/kinesin>; Myosin Home Page, <http://proweb.org/myosin>.
54. Y. Lindqvist and G. Schneider, *Curr. Opin. Struct. Biol.* **7**, 422 (1997).
55. Supported by grants from NIH (GM29009) and U.S. Department of Energy (DE-FG03-97ER62382). S.P. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. T.K.A. is a Royal Society University Research Fellow. P.B. thanks J. Schultz and M. Huynen for helpful discussions.